On Boundedness of Q-Learning Iterates for Stochastic Shortest Path Problems

Citation: Yu, Huizhen, and Dimitri P. Bertsekas. "On Boundedness of Q-Learning Iterates for Stochastic Shortest Path Problems." Mathematics of Operations Research 38, no. 2 (May 2013): 209–227.

As Published: http://dx.doi.org/10.1287/moor.1120.0562

Publisher: Institute for Operations Research and the Management Sciences (INFORMS)

Persistent URL: http://hdl.handle.net/1721.1/93744

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

On Boundedness of Q-Learning Iterates for Stochastic Shortest Path Problems

Huizhen Yu* Dimitri P. Bertsekas†

Abstract

We consider a totally asynchronous stochastic approximation algorithm, Q-learning, for solving finite space stochastic shortest path (SSP) problems, which are total cost Markov decision processes with an absorbing and cost-free state. For the most commonly used SSP models, existing convergence proofs assume that the sequence of Q-learning iterates is bounded with probability one, or some other condition that guarantees boundedness. We prove that the sequence of iterates is naturally bounded with probability one, thus furnishing the boundedness condition in the convergence proof by Tsitsiklis [Tsi94] and establishing completely the convergence of Q-learning for these SSP models.

*Laboratory for Information and Decision Systems (LIDS), M.I.T. (janey_yu@mit.edu).
†Laboratory for Information and Decision Systems (LIDS) and Dept. EECS, M.I.T. (dimitrib@mit.edu).

1 Introduction

Stochastic shortest path (SSP) problems are Markov decision processes (MDP) in which there exists an absorbing and cost-free state, and the goal is to reach that state with minimal expected cost. In this paper we focus on finite state-and-control models under the undiscounted total cost criterion. We call a policy *proper* if under that policy the goal state is reached with probability 1 (w.p.1) for every initial state, and *improper* otherwise. Let $\Pi_{\rm SD}$ denote the set of stationary and deterministic policies. We consider a broad class of SSP models, which satisfy the following general assumption introduced in Bertsekas and Tsitsiklis [BT91]:

Assumption 1.1.

(i) There is at least one proper policy in $\Pi_{\rm SD}$*, and*
(ii) any improper policy in $\Pi_{\rm SD}$ *incurs infinite cost for at least one initial state.*

We will analyze a totally asynchronous stochastic approximation algorithm, the Q-learning algorithm (Watkins [Wat89], Tsitsiklis [Tsi94]), for solving SSP problems. This algorithm generates a sequence of so-called Q-factors $\{Q_t\}$, which represent expected costs associated with initial state-control pairs, and it aims to obtain in the limit the optimal Q-factors Q^* of the problem, from which the optimal costs and optimal policies can be determined.

Under Assumption 1.1, Tsitsiklis [Tsi94, Theorem 2 and Theorem 4(c)] proved that if the sequence $\{Q_t\}$ of Q-learning iterates is bounded w.p.1, then $\{Q_t\}$ converges to the optimal Q-factors Q^* w.p.1. Regarding the boundedness condition, earlier results given in [Tsi94, Lemma 9] and the book by Bertsekas and Tsitsiklis [BT96, Sec. 5.6] show that it is satisfied in the special case where both the one-stage costs and the initial values Q_0 are nonnegative. Alternative to [Tsi94], there is also a line of convergence analysis of Q-learning given in Abounadi, Bertsekas and Borkar [ABB02], which does not require the boundedness condition. However, it requires a more restrictive asynchronous computation framework than the totally asynchronous framework treated in [Tsi94]; in particular, it requires some additional conditions on the timing and frequency of component updates in Q-learning.

In this paper we prove that $\{Q_t\}$ is naturally bounded w.p.1 for SSP models satisfying Assumption 1.1. Our result thus furnishes the boundedness condition in the convergence proof by Tsitsiklis [Tsi94] and, together with the latter, establishes completely the convergence of Q-learning for these SSP models.

This boundedness result is useful as well in other contexts concerning SSP problems. In particular, it is used in the convergence analysis of a new Q-learning algorithm for SSP, proposed recently by the authors [YB11], where the boundedness of the iterates of the new algorithm was related to that of the classical Q-learning algorithm considered here. The line of analysis developed in this paper has also been applied by Yu in [Yu11b] to show the boundedness and convergence of Q-learning for stochastic games of the SSP type.

We organize the paper and the results as follows. In Section 2 we introduce notation and preliminaries. In Section 3 we give the boundedness proof. First we show in Section 3.1 that $\{Q_t\}$ is bounded above w.p.1. We then give in Section 3.2 a short proof that $\{Q_t\}$ is bounded below w.p.1 for a special case with nonnegative expected one-stage costs. In Section 3.3 we prove that $\{Q_t\}$ is bounded below w.p.1 for the general case; the proof is long, so we divide it in several steps given in separate subsections. In Section 4 we illustrate some of these proof steps using a simple example.

2 Preliminaries

2.1 Notation and Definitions

Let $S_o = \{0, 1, \ldots, n\}$ denote the state space, where state 0 is the absorbing and cost-free goal state. Let $S = S_o \setminus \{0\}$. For each state $i \in S$, let $U(i)$ denote the finite set of feasible controls, and for notational convenience, let $U(0) = \{0\}$. We denote by $\mathcal{U}$ the control space, $\mathcal{U} = \cup_{i \in S_o} U(i)$. We define R_o to be the set of state and feasible control pairs, i.e., $R_o = \{(i,u), i \in S_o, u \in U(i)\}$, and we define $R = R_o \setminus \{(0,0)\}$.

The state transitions and associated one-stage costs are defined as follows. From state i with control $u \in U(i)$, a transition to state j occurs with probability $p_{ij}(u)$ and incurs a one-stage cost $\hat{g}(i,u,j)$ or more generally, a random one-stage cost $\hat{g}(i,u,j,\omega)$ where ω is a random disturbance. In the latter case random one-stage costs are all assumed to have finite variance. Let the expected one-stage cost of applying control u at state i be $g(i,u)$. For state 0, $p_{00}(0) = 1$ and the self transition incurs cost 0.

We denote a general history-dependent, randomized policy by π. A randomized Markov policy is a policy of the form $\pi = \{\nu_0, \nu_1, \ldots\}$, where each function ν_k, $k \geq 0$, maps each state $i \in S_o$ to a probability distribution $\nu_k(\cdot \mid i)$ over the set of feasible controls $U(i)$. A randomized Markov policy of the form $\{\nu, \nu, \ldots\}$ is said to be a stationary randomized policy and is also denoted by ν. A stationary deterministic policy is a stationary randomized policy that for each state i assigns probability 1 to a single control $\mu(i)$ in $U(i)$; the policy is also denoted by μ.

The problem is to solve the total cost MDP on S_o, where we define the total cost of a policy π for initial state $i \in S$ to be

$$J^\pi(i) = \liminf_{k \to \infty} J_k^\pi(i),$$

with $J_k^\pi(i)$ being the expected k-stage cost of π starting from state i. Assumption 1.1 is stated for this total cost definition. Under Assumption 1.1, it is established in [BT91] that the Bellman equation (or the total cost optimality equation)

$$J(i) = (TJ)(i) \stackrel{\text{def}}{=} \min_{u \in U(i)} \Big\{ g(i,u) + \sum_{j \in S} p_{ij}(u)\, J(j) \Big\}, \qquad i \in S, \tag{2.1}$$

has a unique solution, which is the optimal cost function J^*, and there exists an optimal policy in Π_{SD}, which is proper of course.

The Q-learning algorithm operates on the so-called Q-factors, $Q = \{Q(i,u), (i,u) \in R_o\} \in \Re^{|R_o|}$. They represent costs associated with initial state-control pairs. For each state-control pair $(i,u) \in R_o$, the optimal Q-factor $Q^*(i,u)$ is the cost of starting from state i, applying control u, and afterwards following an optimal policy. (Here $Q^*(0,0) = 0$ of course.) Then, by the results of [BT91] mentioned above, under Assumption 1.1, the optimal Q-factors and optimal costs are related by

$$Q^*(i,u) = g(i,u) + \sum_{j \in S} p_{ij}(u)\, J^*(j), \qquad J^*(i) = \min_{u \in U(i)} Q^*(i,u), \qquad (i,u) \in R,$$

and Q^* restricted to R is the unique solution of the Bellman equation for Q-factors:

$$Q(i,u) = (FQ)(i,u) \stackrel{\text{def}}{=} g(i,u) + \sum_{j \in S} p_{ij}(u) \min_{v \in U(j)} Q(j,v), \qquad (i,u) \in R. \tag{2.2}$$

Under Assumption 1.1, the Bellman operators T and F given in Eqs. (2.1), (2.2) are not necessarily contraction mappings with respect to the sup-norm $\|\cdot\|_\infty$, but are only nonexpansive. They would be contractions with respect a weighted sup-norm if all policies were proper (see [BT96, Prop. 2.2,

p. 23-24]), and the convergence of Q-learning in that case was established by Tsitsiklis [Tsi94, Theorem 3 and Theorem 4(b)]. A fact that will be used later in our analysis is that for a proper policy $\mu \in \Pi_{\rm SD}$, the associated Bellman operator F_μ given by

$$(F_\mu Q)(i,u) = g(i,u) + \sum_{j \in S} p_{ij}(u) Q(j, \mu(j)), \qquad (i,u) \in R, \tag{2.3}$$

is a weighted sup-norm contraction, with the norm and the modulus of contraction depending on μ. This fact also follows from [BT96, Prop. 2.2, p. 23-24].

2.2 Q-Learning Algorithm

The Q-learning algorithm is an asynchronous stochastic iterative algorithm for finding Q^*. Given an initial $Q_0 \in \Re^{|R_o|}$ with $Q_0(0,0) = 0$, the algorithm generates a sequence $\{Q_t\}$ by updating a subset of Q-factors at each time and keeping the rest unchanged. In particular, $Q_t(0,0) = 0$ for all t. For each $(i,u) \in R$ and $t \geq 0$, let $j_t^{iu} \in S_o$ be the successor state of a random transition from state i after applying control u, generated at time t according to the transition probability $p_{ij}(u)$. Then, with $s = j_t^{iu}$ as a shorthand to simplify notation, the iterate $Q_{t+1}(i,u)$ is given by

$$Q_{t+1}(i,u) = \big(1 - \gamma_t(i,u)\big) Q_t(i,u) + \gamma_t(i,u) \Big(g(i,u) + \omega_t(i,u) + \min_{v \in U(s)} Q_{\tau_t^{sv}(i,u)}(s,v) \Big). \tag{2.4}$$

The variables in the above iteration need to satisfy certain conditions, which will be specified shortly. First we describe what these variables are.

(i) $\gamma_t(i,u) \geq 0$ is a stepsize parameter, and $\gamma_t(i,u) = 0$ if the (i,u)th component is not selected to be updated at time t.

(ii) $g(i,u) + \omega_t(i,u)$ is the random one-stage cost of the transition from state i to j_t^{iu} with control u, i.e., $\omega_t(i,u)$ is the difference between the transition cost and its expected value.

(iii) $\tau_t^{jv}(i,u), (j,v) \in R_o$, are nonnegative integers with $\tau_t^{jv}(i,u) \leq t$. We will refer to them as the delayed times. In a distributed asynchronous computation model, if we associate a processor with each component (i,u), whose task is to update the Q-factor for (i,u), then $t - \tau_t^{jv}(i,u)$ can be viewed as the "communication delay" between the processors at (i,u) and (j,v) at time t.

We now describe the conditions on the variables. We regard all the variables in the Q-learning algorithm as random variables on a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$. This means that the stepsizes and delayed times can be chosen based on the history of the algorithm. To determine the values of these variables, including which components to update at time t, the algorithm may use auxiliary variables that do not appear in Eq. (2.4). Thus, to describe rigorously the dependence relation between the variables, it is convenient to introduce a family $\{\mathcal{F}_t\}$ of increasing sub-σ-fields of $\mathcal{F}$. Then the following *information structure condition* is required: Q_0 is $\mathcal{F}_0$-measurable, and

> for every (i,u) and $(j,v) \in R$ and $t \geq 0$, $\gamma_t(i,u)$ and $\tau_t^{jv}(i,u)$ are $\mathcal{F}_t$-measurable,
> and $\omega_t(i,u)$ and j_t^{iu} are $\mathcal{F}_{t+1}$-measurable.

The condition means that in iteration (2.4), the algorithm either chooses the stepsize $\gamma_t(i,u)$ and the delayed times $\tau_t^{jv}(i,u), (j,v) \in R$, before generating j_t^{iu}, or it chooses the values of the former variables in a way that does not use the information of j_t^{iu}. We note that although this condition seems abstract, it is naturally satisfied by the algorithm in practice.

In probabilistic terms and with the notation just introduced, the successor states and random transition costs appearing in the algorithm need to satisfy the following relations: for all $(i,u) \in R$ and $t \geq 0$,

$$\mathbf{P}(j_t^{iu} = j \mid \mathcal{F}_t) = p_{ij}(u), \qquad \forall\, j \in S_o, \tag{2.5}$$

$$\mathrm{E}\big[\omega_t(i,u) \mid \mathcal{F}_t\big] = 0, \qquad \mathrm{E}\big[\omega_t^2(i,u) \mid \mathcal{F}_t\big] \leq C, \tag{2.6}$$

where C is some deterministic constant. There are two more conditions on the algorithm. The totally asynchronous computation framework has the following minimal requirement on the delayed times used in each component update: w.p.1,

$$\lim_{t\to\infty} \tau_t^{jv}(i,u) = \infty, \qquad \forall\,(i,u),(j,v) \in R. \tag{2.7}$$

As a stochastic approximation algorithm, the standard stepsize condition is required: w.p.1,

$$\sum_{t\geq 0} \gamma_t(i,u) = \infty, \qquad \sum_{t\geq 0} \gamma_t(i,u)^2 < \infty, \qquad \forall\,(i,u) \in R. \tag{2.8}$$

We collect the algorithmic conditions mentioned above in one assumption below. We note that these conditions are natural and fairly mild for the Q-learning algorithm.

Assumption 2.1 (Algorithmic conditions)**.** *The information structure condition holds, and w.p.*1*, Eqs. (2.5)-(2.8) are satisfied.*

For boundedness of the Q-learning iterates, the condition (2.7) is in fact not needed, (which is not surprising intuitively, since bounded delayed times cannot contribute to instability of the iterates). We therefore state also a weaker version of Assumption 2.1, excluding condition (2.7), and we will use it later in the boundedness results for the algorithm.

Assumption 2.2. *The information structure condition holds, and w.p.*1*, Eqs. (2.5), (2.6), (2.8) are satisfied.*

2.3 Convergence of Q-Learning: Earlier Results

The following convergence and boundedness results for Q-learning in SSP problems are established essentially in [Tsi94]; see also [BT96, Sections 4.3 and 5.6].

Theorem 2.1 ([Tsi94])**.** *Let $\{Q_t\}$ be the sequence generated by the iteration (2.4) with any given initial Q_0. Then, under Assumption 2.1, $\{Q_t\}$ converges to Q^* w.p.*1 *if either of the following holds:*

(i) all policies of the SSP are proper;

*(ii) the SSP satisfies Assumption 1.1 and in addition, $\{Q_t\}$ is bounded w.p.*1*.*

*In case (i), we also have that $\{Q_t\}$ is bounded w.p.*1 *under Assumption 2.2 (instead of Assumption 2.1).*

Note that for a proper policy $\mu \in \Pi_{\mathrm{SD}}$, by considering the SSP problem that has μ as its only policy, the conclusions of Theorem 2.1 in case (i) apply also to the evaluation of policy μ with Q-learning. In this context, Q^* in the conclusions corresponds to the Q-factor vector Q_μ, which is the unique fixed point of the weighted sup-norm contraction mapping F_μ (see Eq. (2.3)).

The contribution of this paper is to remove the boundedness requirement on $\{Q_t\}$ in case (ii). Our proof arguments will be largely different from those used to establish the preceding theorem. For completeness, however, in the rest of this section, we explain briefly the basis of the analysis that gives Theorem 2.1, and the conditions involved.

In the analytical framework of [Tsi94], we view iteration (2.4) as a stochastic approximation algorithm and rewrite it equivalently as

$$Q_{t+1}(i,u) = \big(1-\gamma_t(i,u)\big)Q_t(i,u) + \gamma_t(i,u)\big(FQ_t^{(iu)}\big)(i,u) + \gamma_t(i,u)\,\tilde{\omega}_t(i,u), \tag{2.9}$$

where F is the Bellman operator given by Eq. (2.2); $Q_t^{(iu)}$ denotes the vector of Q-factors with components $Q_{\tau_t^{jv}(i,u)}(j,v), (j,v) \in R_o$, (which involve the delayed times); and $\tilde{\omega}_t(i,u)$ is a noise term given by

$$\tilde{\omega}_t(i,u) = g(i,u) + \omega_t(i,u) + \min_{v\in U(s)} Q_{\tau_t^{sv}(i,u)}(s,v) - \big(FQ_t^{(iu)}\big)(i,u)$$

(with $s = j_t^{iu}$). The noise terms $\tilde{\omega}_t(i,u), (i,u) \in R$, are $\mathcal{F}_{t+1}$-measurable. Conditional on $\mathcal{F}_t$, they can be shown to have zero mean and meet a requirement on the growth of the conditional variance, when the Q-learning algorithm satisfies certain conditions (the same as those in Assumption 2.1 except for a slightly stronger stepsize condition, which will be explained shortly). We then analyze iteration (2.9) as a special case of an asynchronous stochastic approximation algorithm where F is either a contraction or a monotone nonexpansive mapping (with respect to the sup-norm) and Q^* is the unique fixed point of F. These two cases of F correspond to the two different SSP model assumptions in Theorem 2.1: when all policies of the SSP are proper, F is a weighted sup-norm contraction, whereas when Assumption 1.1 holds, F is monotone and nonexpansive (see Section 2.1). The conclusions of Theorem 2.1 for case (i) follow essentially from [Tsi94, Theorems 1 and 3] for contraction mappings, whereas Theorem 2.1 in case (ii) follows essentially from [Tsi94, Theorem 2] for monotone nonexpansive mappings.

A specific technical detail relating to the stepsize condition is worth mentioning. To apply the results of [Tsi94] here, we first consider, without loss of generality, the case where all stepsizes are bounded by some deterministic constant. Theorem 2.1 under this additional condition then follows directly from [Tsi94]; see also [BT96, Section 4.3].[1] (We mention that the technical use of this additional stepsize condition is only to ensure that the noise terms $\tilde{\omega}_t(i,u), (i,u) \in R$, have well-defined conditional expectations.) We then remove the additional stepsize condition and obtain Theorem 2.1 as the immediate consequence, by using a standard, simple truncation technique as follows. For each positive integer m, define truncated stepsizes

$$\hat{\gamma}_t^m(i,u) = \min\{m, \gamma_t(i,u)\}, \qquad (i,u) \in R,$$

which are by definition bounded by m, and consider the sequence $\{\hat{Q}_t^m\}$ generated by iteration (2.4) with $\hat{Q}_0^m = Q_0$ and with $\hat{\gamma}_t^m(i,u)$ in place of $\gamma_t(i,u)$. This sequence has the following properties. If the original sequence $\{Q_t\}$ satisfies Assumption 2.1 or 2.2, then so does $\{\hat{Q}_t^m\}$. Moreover, since the original stepsizes $\gamma_t(i,u), t \geq 0, (i,u) \in R$, are bounded w.p.1, we have that for each sample path from a set of probability one, $\{Q_t\}$ coincides with $\{\hat{Q}_t^m\}$ for some sufficiently large integer m. The latter means that if for each m, $\{\hat{Q}_t^m\}$ converges to Q^* (or $\{\hat{Q}_t^m\}$ is bounded) w.p.1, then the same holds for $\{Q_t\}$. Hence the conclusions of Theorem 2.1 for case (i) are direct consequences of applying the weaker version of the theorem mentioned earlier to the sequences $\{\hat{Q}_t^m\}$ for each m. Case (ii) of Theorem 2.1 follows from exactly the same argument, in view of the fact that under Assumption 2.1, if $\{Q_t\}$ is bounded w.p.1, then $\{\hat{Q}_t^m\}$ is also bounded w.p.1 for each m. [To see this, observe that by condition (2.8), the stepsizes in $\{Q_t\}$ and $\{\hat{Q}_t^m\}$ coincide for t sufficiently large; more precisely, w.p.1, there exists some finite (path-dependent) time $\bar{t}$ such that for all $t \geq \bar{t}$ and

[1] The stepsize condition appearing in [Tsi94] is slightly different than condition (2.8); it is $\sum_{t\geq 0}\gamma_t(i,u)^2 < C$ w.p.1, for some (deterministic) constant C, instead of C being ∞, and in addition, it is required that $\gamma_t(i,u) \in [0,1]$. However, by strengthening one technical lemma (Lemma 1) in [Tsi94] so that its conclusions hold under the weaker condition (2.8), the proof of [Tsi94] is essentially intact under the latter condition. The details of the analysis can be found in [BT96, Prop. 4.1 and Example 4.3, p. 141-143] (see also Cor. 4.1 and Section 4.3.6 therein). A reproduction of the proofs in [Tsi94, BT96] with slight modifications is also available [Yu11a].

$(i,u) \in R$, $\hat{\gamma}_t^m(i,u) = \gamma_t(i,u) \in [0,1]$. It then follows by the definition of $\{\hat{Q}_t^m\}$ that $\|Q_t - \hat{Q}_t^m\|_\infty \leq \max_{\tau \leq \bar{t}} \|Q_\tau - \hat{Q}_\tau^m\|_\infty$ for all $t \geq \bar{t}$.] So, technically speaking, Theorem 2.1 with the general stepsizes is a corollary of its weaker version mentioned earlier.

3 Main Results

We will prove in this section the following theorem. It furnishes the boundedness condition required in [Tsi94, Theorem 2] (see Theorem 2.1(ii)), and together with the latter, establishes completely the convergence of $\{Q_t\}$ to Q^* w.p.1.

Theorem 3.1. *Under Assumptions 1.1 and 2.2, for any given initial Q_0, the sequence $\{Q_t\}$ generated by the Q-learning iteration (2.4) is bounded w.p.*1.

Our proof consists of several steps which will be given in separate subsections. First we show that $\{Q_t\}$ is bounded above w.p.1. This proof is short and uses the contraction property of the Bellman operator F_μ associated with a proper policy μ in Π_{SD}. A similar idea has been used in earlier works [Tsi94, Lemma 9] and [BT96, Prop. 5.6, p. 249] to prove the boundedness of iterates for certain nonnegative SSP models.

In the proofs of this section, for brevity, we will partially suppress the word "w.p.1" when the algorithmic conditions are concerned. Whenever a subset of sample paths with a certain property is considered, it will be implicitly assumed to be the intersection of the set of paths with that property and the set of paths that satisfy the assumption on the algorithm currently in effect (e.g., Assumption 2.2 or 2.1). In the proofs, the notation "$\overset{a.s.}{\to}$" stands for almost sure convergence.

3.1 Boundedness from Above

Proposition 3.1. *Under Assumptions 1.1(i) and 2.2, for any given initial Q_0, the sequence $\{Q_t\}$ generated by the Q-learning iteration (2.4) is bounded above w.p.*1.

Proof. Let μ be any proper policy in $\Pi_{\rm SD}$, which exists by Assumption 1.1(i). First we define iterates (random variables) $\{\hat{Q}_t\}$ on the same probability space as the Q-learning iterates $\{Q_t\}$. Let $\hat{Q}_0 = Q_0$ and $\hat{Q}_t(0,0) = 0$ for $t \geq 0$. For each $(i,u) \in R$ and $t \geq 0$, let

$$\hat{Q}_{t+1}(i,u) = \big(1 - \gamma_t(i,u)\big)\hat{Q}_t(i,u) + \gamma_t(i,u)\Big(g(i,u) + \omega_t(i,u) + \hat{Q}_{\tau_t^{s\bar{v}}(i,u)}\big(j_t^{iu}, \mu(j_t^{iu})\big)\Big),$$

where in the superscript of $\tau_t^{s\bar{v}}(i,u)$, s is a shorthand for j_t^{iu} and $\bar{v}$ is a shorthand for $\mu(j_t^{iu})$, introduced to avoid notational clutter; and $\gamma_t(i,u)$, j_t^{iu} and $\omega_t(i,u)$, as well as the delayed times $\tau_t^{jv}(i,u), (j,v) \in R_o$, are the same random variables that appear in the Q-learning algorithm (2.4).

The sequence $\{\hat{Q}_t\}$ is generated by the Q-learning algorithm (2.4) for the SSP problem that has the proper policy μ as its only policy, and involves the mapping F_μ, which is a weighted sup-norm contraction (see Section 2.1 and the discussion following Theorem 2.1). The sequence $\{\hat{Q}_t\}$ also satisfies Assumption 2.2 (since $\{\hat{Q}_t\}$ and $\{Q_t\}$ involve the same stepsizes, transition costs and delayed times). Therefore, by Theorem 2.1(i), $\{\hat{Q}_t\}$ is bounded w.p.1.

Consider now any sample path from the set of probability one on which $\{\hat{Q}_t\}$ is bounded. In view of the stepsize condition (2.8), there exists a time $\bar{t}$ such that $\gamma_t(i,u) \leq 1$ for all $t \geq \bar{t}$ and $(i,u) \in R$. Let $\Delta = \max_{\tau \leq \bar{t}} \max_{(i,u) \in R} \big(Q_\tau(i,u) - \hat{Q}_\tau(i,u)\big)$. Then

$$Q_\tau(i,u) \leq \hat{Q}_\tau(i,u) + \Delta, \qquad \forall\, (i,u) \in R,\ \tau \leq \bar{t}.$$

We show by induction that this relation also holds for all $\tau > \bar{t}$. To this end, suppose that for some $t \geq \bar{t}$, the relation holds for all $\tau \leq t$. Then, for each $(i,u) \in R$, we have that

$$\begin{aligned} Q_{t+1}(i,u) &\leq \big(1-\gamma_t(i,u)\big)Q_t(i,u) + \gamma_t(i,u)\Big(g(i,u) + \omega_t(i,u) + Q_{\tau_t^{s\bar{v}}(i,u)}(s,\bar{v})\Big) \\ &\leq \big(1-\gamma_t(i,u)\big)\big(\hat{Q}_t(i,u)+\Delta\big) + \gamma_t(i,u)\Big(g(i,u) + \omega_t(i,u) + \hat{Q}_{\tau_t^{s\bar{v}}(i,u)}\big(s,\bar{v}\big) + \Delta\Big) \\ &= \hat{Q}_{t+1}(i,u) + \Delta, \end{aligned}$$

where the first inequality follows from the definition of Q_{t+1} and the fact that $\gamma_t(i,u) \geq 0$, the second inequality follows from the induction hypothesis and the fact that $\gamma_t(i,u) \in [0,1]$, and the last equality follows from the the definition of $\hat{Q}_{t+1}$. This completes the induction and shows that $\{Q_t\}$ is bounded above w.p.1. □

3.2 Boundedness from Below for a Special Case

The proof that $\{Q_t\}$ is bounded below w.p.1 is long and consists of several steps to be given in the next subsection. For a special case with nonnegative expected one-stage costs, there is a short proof, which we give here. Together with Prop. 3.1, it provides a short proof of the boundedness and hence convergence of the Q-learning iterates for a class of nonnegative SSP models satisfying Assumption 1.1. Earlier works [Tsi94, Lemma 9] and [BT96, Prop. 5.6, p. 249] have also considered nonnegative SSP models and established convergence results for them, but under stronger assumptions than ours. [In particular, it is assumed there that all transitions incur costs $\hat{g}(i,u,j,\omega) \geq 0$, as well as other conditions, so that all iterates are nonnegative.] To keep the proof simple, we will use Assumption 2.1, although Assumption 2.2 would also suffice.

Proposition 3.2. *Suppose that $g(i,u) \geq 0$ for all $(i,u) \in R$ and moreover, for those (i,u) with $g(i,u) = 0$, every possible transition from state i under control u incurs cost 0. Then, under Assumption 2.1, for any given initial Q_0, the sequence $\{Q_t\}$ generated by the Q-learning iteration (2.4) is bounded below w.p.1.*

Proof. We write $\{Q_t\}$ as the sum of two processes: for each $(i,u) \in R_o$,

$$Q_t(i,u) = \tilde{g}_t(i,u) + Y_t(i,u), \qquad t \geq 0, \tag{3.1}$$

where $\tilde{g}_t(0,0) = g(0,0) = 0$ and $Y_t(0,0) = 0$ for all t, and for each $(i,u) \in R$,

$$\begin{aligned} \tilde{g}_{t+1}(i,u) &= \big(1-\gamma_t(i,u)\big)\tilde{g}_t(i,u) + \gamma_t(i,u)\big(g(i,u)+\omega_t(i,u)\big), \\ Y_{t+1}(i,u) &= \big(1-\gamma_t(i,u)\big)Y_t(i,u) + \gamma_t(i,u)\min_{v\in U(s)} Q_{\tau_t^{sv}(i,u)}(s,v), \end{aligned}$$

with $\tilde{g}_0 \equiv 0$, $Y_0 = Q_0$, and s being a shorthand for j_t^{iu} (to avoid notational clutter). Using the conditions (2.6) and (2.8) of the Q-learning algorithm, it follows from the standard theory of stochastic approximation (see e.g., [BT96, Prop. 4.1 and Example 4.3, p. 141-143] or [KY03, Bor08]) that $\tilde{g}_t(i,u) \overset{a.s.}{\to} g(i,u)$ for all $(i,u) \in R$.[2]

Consider any sample path from the set of probability one, on which this convergence takes place. Then by Eq. (3.1), on that sample path, $\{Q_t\}$ is bounded below if and only if $\{Y_t\}$ is bounded below. Now from the definition of Y_t and Eq. (3.1) we have

$$Y_{t+1}(i,u) = \big(1-\gamma_t(i,u)\big)Y_t(i,u) + \gamma_t(i,u)\min_{v\in U(s)}\big(\tilde{g}_{\tau_t^{sv}(i,u)}(s,v) + Y_{\tau_t^{sv}(i,u)}(s,v)\big). \tag{3.2}$$

[2] This convergence follows from a basic result of stochastic approximation theory (see the aforementioned references) if besides (2.6) and (2.8), it is assumed in addition that the stepsizes are bounded by some (deterministic) constant. The desired result then follows by removing the additional condition with the stepsize truncation proof technique described in Section 2.3. More details can also be found in [Yu11a]; Lemma 1 therein implies the convergence desired here.

By condition (2.7) of the Q-learning algorithm, and in view also of our assumption on one-stage costs, the convergence $\tilde{g}_t(j,v) \overset{a.s.}{\to} g(j,v)$ for all $(j,v) \in R$ implies that on the sample path under our consideration, for all t sufficiently large,

$$\tilde{g}_{\tau_t^{jv}(i,u)}(j,v) \geq 0, \qquad \forall\, (j,v) \in R_o.$$

Therefore, using Eq. (3.2) and the fact that eventually $\gamma_t(i,u) \in [0,1]$ [cf. Eq. (2.8)], we have that for all t sufficiently large and for all $(i,u) \in R$,

$$\begin{aligned} Y_{t+1}(i,u) &\geq \big(1-\gamma_t(i,u)\big)Y_t(i,u) + \gamma_t(i,u) \min_{v\in U(s)} Y_{\tau_t^{sv}(i,u)}(s,v) \\ &\geq \min_{\tau\leq t} \min_{(j,v)\in R_o} Y_\tau(j,v), \end{aligned}$$

which implies that for all t sufficiently large,

$$\min_{\tau\leq t+1} \min_{(j,v)\in R_o} Y_t(j,v) \geq \min_{\tau\leq t} \min_{(j,v)\in R_o} Y_\tau(j,v).$$

Hence $\{Y_t\}$ is bounded below on that sample path. The proof is complete. □

3.3 Boundedness from Below in General

In this section, we will prove the following result in several steps. Together with Prop. 3.1 it implies Theorem 3.1.

Proposition 3.3. *Under Assumptions 1.1 and 2.2, the sequence $\{Q_t\}$ generated by the Q-learning iteration (2.4) is bounded below w.p.*1.

The proof can be outlined roughly as follows. In Section 3.3.1 we will introduce an auxiliary sequence $\{\tilde{Q}_t\}$ of a certain form such that $\{\tilde{Q}_t\}$ is bounded below w.p.1 if and only if $\{Q_t\}$ is bounded below w.p.1. In Sections 3.3.2 and 3.3.3 we will give, for any given $\delta > 0$, a specific construction of the sequence $\{\tilde{Q}_t\}$ for each sample path from a set of probability 1, such that each $\tilde{Q}_t(i,u)$ can be interpreted as the expected total cost of some randomized Markov policy for a time-inhomogeneous SSP problem that can be viewed as a "δ-perturbation" of the original problem. Finally, to complete the proof, we will show in Section 3.3.4 that when δ is sufficiently small, the expected total costs achievable in any of these "perturbed" SSP problems can be bounded uniformly from below, so that the auxiliary sequence $\{\tilde{Q}_t\}$ constructed for the corresponding δ must be bounded below w.p.1. This then implies that the Q-learning iterates $\{Q_t\}$ must be bounded below w.p.1.

In what follows, let Ω' denote the set of sample paths on which the algorithmic conditions in Assumption 2.2 hold. Note that Ω' has probability one under Assumption 2.2.

3.3.1 Auxiliary sequence $\{\tilde{Q}_t\}$

The first step of our proof is a technically important observation. Let us write the Q-learning iterates given in Eq. (2.4) equivalently, for all $(i,u) \in R$ and $t \geq 0$, as

$$Q_{t+1}(i,u) = \big(1-\gamma_t(i,u)\big)Q_t(i,u) + \gamma_t(i,u)\big(g(i,u) + \omega_t(i,u) + Q_{\tau_t^{sv}(i,u)}(j_t^{iu}, v_t^{iu})\big), \tag{3.3}$$

where v_t^{iu} is a control that satisfies

$$v_t^{iu} \in \arg\min_{\bar{v}\in U(s)} Q_{\tau_t^{s\bar{v}}(i,u)}(j_t^{iu}, \bar{v}), \tag{3.4}$$

and s, v in the superscript of $\tau_t^{sv}(i,u)$ are shorthand notation: s stands for the state j_t^{iu} and v now stands for the control v_t^{iu}. We observe the following. Suppose we define an auxiliary sequence $\{\tilde{Q}_t\}$ where

$$\tilde{Q}_t(0,0) = 0, \qquad t \geq 0, \tag{3.5}$$

and for some nonnegative integer t_0 and for all $(i,u) \in R$,

$$\tilde{Q}_{t+1}(i,u) = \big(1-\gamma_t(i,u)\big)\tilde{Q}_t(i,u) + \gamma_t(i,u)\big(g(i,u) + \omega_t(i,u) + \tilde{Q}_{\tau_t^{sv}(i,u)}(j_t^{iu}, v_t^{iu})\big), \qquad t \geq t_0, \tag{3.6}$$

$$\tilde{Q}_t(i,u) = \tilde{Q}_{t_0}(i,u), \qquad t \leq t_0. \tag{3.7}$$

Let us consider each sample path from the set Ω'. In view of Eq. (2.8), there exists $t_0' \geq t_0$ such that $\gamma_t(i,u) \in [0,1]$ for all $t \geq t_0'$ and $(i,u) \in R$. By Eqs. (3.3) and (3.6), we then have that for all $t \geq t_0'$ and $(i,u) \in R$,

$$\begin{aligned}
\big|Q_{t+1}(i,u) - \tilde{Q}_{t+1}(i,u)\big| &\leq \big(1-\gamma_t(i,u)\big)\big|Q_t(i,u) - \tilde{Q}_t(i,u)\big| \\
&\quad + \gamma_t(i,u)\big|Q_{\tau_t^{sv}(i,u)}(j_t^{iu}, v_t^{iu}) - \tilde{Q}_{\tau_t^{sv}(i,u)}(j_t^{iu}, v_t^{iu})\big| \\
&\leq \max_{\tau \leq t} \|Q_\tau - \tilde{Q}_\tau\|_\infty,
\end{aligned}$$

which implies

$$\max_{\tau \leq t+1} \|Q_\tau - \tilde{Q}_\tau\|_\infty \leq \max_{\tau \leq t} \|Q_\tau - \tilde{Q}_\tau\|_\infty. \tag{3.8}$$

Therefore, on that sample path, $\{Q_t\}$ is bounded below if and only if $\{\tilde{Q}_t\}$ is bounded below. We state this as a lemma.

Lemma 3.1. *For any sample path from the set Ω', and for any values of t_0 and $\tilde{Q}_{t_0}$, the Q-learning sequence $\{Q_t\}$ is bounded below if and only if $\{\tilde{Q}_t\}$ given by Eqs. (3.5)-(3.7) is bounded below.*

This observation is the starting point for the proof of the lower boundedness of $\{Q_t\}$. We will construct a sequence $\{\tilde{Q}_t\}$ that is easier to analyze than $\{Q_t\}$ itself. In particular, we will choose, for each sample path from a set of probability one, the time t_0 and the initial $\tilde{Q}_{t_0}$ in such a way that the auxiliary sequence $\{\tilde{Q}_t\}$ is endowed with a special interpretation and structure relating to perturbed versions of the SSP problem.

3.3.2 Choosing t_0 and initial $\tilde{Q}_{t_0}$ for a sample path

First we introduce some notation and definitions to be used throughout the rest of the proof. For a finite set D, let $\mathcal{P}(D)$ denote the set of probability distributions on D. For $p \in \mathcal{P}(D)$ and $x \in D$, let $p(x)$ denote the probability of x and $\mathrm{supp}(p)$ denote the support of p, $\{x \in D \mid p(x) \neq 0\}$. For $p_1, p_2 \in \mathcal{P}(D)$, we write $p_1 \ll p_2$ if p_1 is absolutely continuous with respect to p_2, that is, $\mathrm{supp}(p_1) \subset \mathrm{supp}(p_2)$. For signed measures p on D, we define the notation $p(x)$ and $\mathrm{supp}(p)$ as well as the notion of absolute continuity similarly. We denote by $\overline{\mathcal{P}}(D)$ the set of signed measures p on D such that $\sum_{x \in D} p(x) = 1$. This set contains the set $\mathcal{P}(D)$.

For each $(i,u) \in R_o$, we define the following. Let $\mathbf{p}_o^{iu} \in \mathcal{P}(S_o)$ correspond to the transition probabilities at (i,u):

$$\mathbf{p}_o^{iu}(j) = p_{ij}(u), \qquad j \in S_o.$$

For each $\delta > 0$, let $A_\delta(i,u) \subset \mathcal{P}(S_o)$ denote the set of probability distributions that are both in the δ-neighborhood of $\mathbf{p}_o^{iu}$ and absolutely continuous with respect to $\mathbf{p}_o^{iu}$, i.e.,

$$A_\delta(i,u) = \big\{\, \mathbf{d} \in \mathcal{P}(S_o) \,\big|\, |\mathbf{d}(j) - p_{ij}(u)| \leq \delta,\ \forall\, j \in S_o, \text{ and } \mathbf{d} \ll \mathbf{p}_o^{iu} \big\}.$$

(In particular, for $(i,u) = (0,0)$, $\mathbf{p}_o^{00}(0) = 1$ and $A_\delta(0,0) = \big\{\mathbf{p}_o^{00}\big\}$.)

Let g denote the vector of expected one-stage costs, $\{g(i,u),\ (i,u) \in R_o\}$. Define B_δ to be the subset of vectors in the δ-neighborhood of g whose $(0,0)$th component is zero: with $\mathbf{c} = \{c(i,u),\ (i,u) \in R_o\}$,

$$B_\delta = \big\{\, \mathbf{c} \,\big|\, c(0,0) = 0 \text{ and } |c(i,u) - g(i,u)| \le \delta,\ \forall\, (i,u) \in R \big\}.$$

We now describe how we choose t_0 and $\tilde{Q}_{t_0}$ for the auxiliary sequence $\{\tilde{Q}_t\}$ on a certain set of sample paths that has probability one. We start by defining two sequences, a sequence $\{\tilde{g}_t\}$ of one-stage cost vectors[3] and a sequence $\{\mathbf{q}_t\}$ of collections of signed measures in $\overline{\mathcal{P}}(S_o)$. They are random sequences defined on the same probability space as the Q-learning iterates, and they can be related to the empirical one-stage costs and empirical transition frequencies on a sample path. We define the sequence $\{\tilde{g}_t\}$ as follows: for $t \ge 0$,

$$\tilde{g}_{t+1}(i,u) = \big(1 - \gamma_t(i,u)\big)\, \tilde{g}_t(i,u) + \gamma_t(i,u)\, \big(g(i,u) + \omega_t(i,u)\big), \qquad \forall\, (i,u) \in R; \tag{3.9}$$

$$\tilde{g}_0(i,u) = 0, \qquad (i,u) \in R; \quad \text{and} \quad \tilde{g}_t(0,0) = 0, \qquad t \ge 0.$$

We define the sequence $\{\mathbf{q}_t\}$ as follows. It has as many components as the size of the set R of state-control pairs. For each $(i,u) \in R$, define the component sequence $\{\mathbf{q}_t^{iu}\}$ by letting $\mathbf{q}_0^{iu}$ be any given distribution in $\mathcal{P}(S_o)$ with $\mathbf{q}_0^{iu} \ll \mathbf{p}_o^{iu}$, and by letting

$$\mathbf{q}_{t+1}^{iu} = \big(1 - \gamma_t(i,u)\big)\, \mathbf{q}_t^{iu} + \gamma_t(i,u)\, \mathbf{e}_{j_t^{iu}}, \qquad t \ge 0, \tag{3.10}$$

where $\mathbf{e}_j$ denotes the indicator of j: $\mathbf{e}_j \in \mathcal{P}(S_o)$ with $\mathbf{e}_j(j) = 1$ for $j \in S_o$. Since the stepsizes $\gamma_t(i,u)$ may exceed 1, in general $\mathbf{q}_t^{iu} \in \overline{\mathcal{P}}(S_o)$. Since j_t^{iu} is a random successor state of state i after applying control u [cf. condition (2.5)], w.p.1,

$$\mathbf{q}_t^{iu} \ll \mathbf{p}_o^{iu}, \qquad t \ge 0. \tag{3.11}$$

By the standard theory of stochastic approximation (see e.g., [BT96, Prop. 4.1 and Example 4.3, p. 141-143] or [KY03, Bor08]; see also Footnote 2), Eqs. (2.6) and (2.8) imply that

$$\tilde{g}_t(i,u) \overset{a.s.}{\to} g(i,u), \qquad \forall\, (i,u) \in R, \tag{3.12}$$

whereas Eqs. (2.5) and (2.8) imply that

$$\mathbf{q}_t^{iu} \overset{a.s.}{\to} \mathbf{p}_o^{iu}, \qquad \forall\, (i,u) \in R. \tag{3.13}$$

Equations (3.13) and (3.11) together imply that w.p.1, eventually $\mathbf{q}_t^{iu}$ lies in the set $\mathcal{P}(S_o)$ of probability distributions. The following is then evident, in view also of the stepsize condition (2.8).

Lemma 3.2. *Let Assumption 2.2 hold. Consider any sample path from the set of probability one of paths which lie in Ω' and on which the convergence in Eqs. (3.12), (3.13) takes place. Then for any $\delta > 0$, there exists a time t_0 such that*

$$\tilde{g}_t \in B_\delta, \qquad \mathbf{q}_t^{iu} \in A_\delta(i,u), \qquad \gamma_t(i,u) \le 1, \qquad \forall\, (i,u) \in R,\ t \ge t_0. \tag{3.14}$$

In the rest of Section 3.3, let us consider any sample path from the set of probability one given in Lemma 3.2. For any given $\delta > 0$, we choose t_0 given in Lemma 3.2 to be the initial time of the auxiliary sequence $\{\tilde{Q}_t\}$. (Note that t_0 depends on the entire path and hence so does $\{\tilde{Q}_t\}$.)

We now define the initial $\tilde{Q}_{t_0}$. Our definition and the proof that follows will involve a stationary randomized policy ν. Recall that $\nu(u \mid i)$ denotes the probability of applying control u at state i under ν, for $u \in U(i), i \in S_o$; recall also that $\mathcal{U} = \cup_{i \in S_o} U(i)$ is the control space. We now regard

[3]The sequence $\{\tilde{g}_t\}$ also appeared in the proof of Prop. 3.2; for convenience, we repeat the definition here.

$\nu(\cdot \mid i)$ as a distribution in $\mathcal{P}(\mathcal{U})$ with its support contained in the feasible control set $U(i)$, [that is, $\nu(u \mid i) = 0$ if $u \notin U(i)$].

To define $\tilde{Q}_{t_0}$, let ν be a proper randomized stationary policy, which exists under Assumption 1.1(i). We define each component $\tilde{Q}_{t_0}(i,u)$ of $\tilde{Q}_{t_0}$ separately, and we associate with $\tilde{Q}_{t_0}(i,u)$ a time-inhomogeneous Markov chain and time-varying one-stage cost functions as follows. For each $(i,u) \in R$, consider a time-inhomogeneous Markov chain $(i_0,u_0),(i_1,u_1),\ldots$ on the space $S_o \times \mathcal{U}$ with initial state $(i_0,u_0) = (i,u)$, whose probability distribution is denoted $\mathbf{P}^{iu}_{t_0}$ and whose transition probabilities at time $k-1$ are given by: for all $(\bar{i},\bar{u}),(\bar{j},\bar{v}) \in R_o$,

$$\begin{aligned}
\mathbf{P}^{iu}_{t_0}\big(i_1 = \bar{j}, u_1 = \bar{v} \mid i_0 = i, u_0 = u\big) &= \mathbf{q}^{iu}_{t_0}(\bar{j}) \cdot \nu(\bar{v} \mid \bar{j}), \qquad \text{for } k = 1,\\
\mathbf{P}^{iu}_{t_0}\big(i_k = \bar{j}, u_k = \bar{v} \mid i_{k-1} = \bar{i}, u_{k-1} = \bar{u}\big) &= p_{\bar{i}\bar{j}}(\bar{u}) \cdot \nu(\bar{v} \mid \bar{j}), \qquad \text{for } k \geq 2,
\end{aligned}$$

where $\mathbf{P}^{iu}_{t_0}(\cdot \mid \cdot)$ denotes conditional probability. (The transition probabilities at those $(\bar{i},\bar{u}) \notin R_o$ can be defined arbitrarily because regardless of their values, the chain with probability one will never visit such state-control pairs at any time.) For each $(i,u) \in R$, we also define time-varying one-stage cost functions $g^{iu,t_0}_k : R_o \mapsto \Re$, $k \geq 0$, by

$$g^{iu,t_0}_0 = \tilde{g}_{t_0}, \quad \text{for } k = 0, \qquad \text{and} \qquad g^{iu,t_0}_k = g, \quad \text{for } k \geq 1.$$

We extend g^{iu,t_0}_k to $S_o \times \mathcal{U}$ by defining its values outside the domain R_o to be $+\infty$, and we will treat $0 \cdot \infty = 0$. This convention will be followed throughout.

We now define

$$\tilde{Q}_{t_0}(i,u) = \mathrm{E}^{\mathbf{P}^{iu}_{t_0}}\Big[\sum_{k=0}^{\infty} g^{iu,t_0}_k(i_k,u_k)\Big], \qquad \forall\,(i,u) \in R, \tag{3.15}$$

where $\mathrm{E}^{\mathbf{P}^{iu}_{t_0}}$ denotes expectation under $\mathbf{P}^{iu}_{t_0}$. The above expectation is well-defined and finite, and furthermore, the order of summation and expectation can be exchanged, i.e.,

$$\tilde{Q}_{t_0}(i,u) = \sum_{k=0}^{\infty} \mathrm{E}^{\mathbf{P}^{iu}_{t_0}}\Big[g^{iu,t_0}_k(i_k,u_k)\Big].$$

This follows from the fact that under $\mathbf{P}^{iu}_{t_0}$, from time 1 onwards, the process $\{(i_k,u_k), k \geq 1\}$ evolves and incurs costs as in the original SSP problem under the stationary proper policy ν. In particular, since ν is a proper policy, $\sum_{k=0}^{\infty} |g^{iu,t_0}_k(i_k,u_k)|$ is finite almost surely with respect to $\mathbf{P}^{iu}_{t_0}$, and hence the summation $\sum_{k=0}^{\infty} g^{iu,t_0}_k(i_k,u_k)$ is well-defined and also finite $\mathbf{P}^{iu}_{t_0}$-almost surely. Since ν is a stationary proper policy for a finite state SSP, we have that under ν, from any state in S, the expected time of reaching the state 0 is finite, and consequently, $\mathrm{E}^{\mathbf{P}^{iu}_{t_0}}\Big[\sum_{k=0}^{\infty} \big|g^{iu,t_0}_k(i_k,u_k)\big|\Big]$ is also finite. It then follows from the dominated convergence theorem that the two expressions given above for $\tilde{Q}_{t_0}(i,u)$ are indeed equal.

3.3.3 Interpreting $\{\tilde{Q}_t\}$ as costs in certain time-inhomogeneous SSP problems

We now show that with the preceding choice of t_0 and initial $\tilde{Q}_{t_0}$, each component of the iterates $\tilde{Q}_t, t \geq t_0$, is equal to, briefly speaking, the expected total cost of a randomized Markov policy (represented by $\{\nu^{iu,t}_k, k \geq 1\}$ below) in a time-inhomogeneous SSP problem whose parameters (transition probabilities and one-stage costs, represented by $\{p^{iu,t}_k, g^{iu,t}_k, k \geq 0\}$ below) lie in the δ-neighborhood of those of the original problem. While the proof of this result is lengthy, it is mostly a straightforward verification. In the next, final step of our analysis, given in Section 3.3.4, we will, for sufficiently small δ, lower-bound the costs of these time-inhomogeneous SSP problems and thereby lower-bound $\{\tilde{Q}_t\}$.

As in the preceding subsection, for any probability distribution $\mathbf{P}$, we write $\mathbf{P}(\cdot \mid \cdot)$ for conditional probability and $\mathrm{E}^{\mathbf{P}}$ for expectation under $\mathbf{P}$. Recall also that the sets $A_\delta(i,u)$ where $(i,u) \in R_o$, and the set B_δ, defined in the preceding subsection, are subsets contained in the δ-neighborhood of the transition probability parameters and expected one-stage cost parameters of the original SSP problem, respectively.

Lemma 3.3. *Let Assumptions 1.1(i) and 2.2 hold. Consider any sample path from the set of probability one given in Lemma 3.2. For any $\delta > 0$, with t_0 and $\tilde{Q}_{t_0}$ given as in Section 3.3.2 for the chosen δ, the iterates $\tilde{Q}_t(i,u)$ defined by Eqs. (3.5)-(3.7) have the following properties for each $(i,u) \in R$ and $t \geq 0$:*

(a) *$\tilde{Q}_t(i,u)$ can be expressed as*

$$\tilde{Q}_t(i,u) = \mathrm{E}^{\mathbf{P}_t^{iu}}\Big[\sum_{k=0}^{\infty} g_k^{iu,t}(i_k,u_k)\Big] = \sum_{k=0}^{\infty} \mathrm{E}^{\mathbf{P}_t^{iu}}\Big[g_k^{iu,t}(i_k,u_k)\Big]$$

for some probability distribution $\mathbf{P}_t^{iu}$ of a Markov chain $\{(i_k,u_k), k \geq 0\}$ on $S_o \times \mathcal{U}$ and one-stage cost functions $g_k^{iu,t} : R_o \mapsto \Re$, $k \geq 0$ (with $g_k^{iu,t} \equiv +\infty$ on $(S_o \times \mathcal{U}) \setminus R_o$).

(b) *The Markov chain $\{(i_k,u_k), k \geq 0\}$ in (a) starts from state $(i_0,u_0) = (i,u)$ and is time-inhomogeneous. Its transition probabilities have the following product form: for all $(\bar{i},\bar{u}), (\bar{j},\bar{v}) \in R_o$,*

$$\begin{aligned}
\mathbf{P}_t^{iu}\big(i_1 = \bar{j}, u_1 = \bar{v} \mid i_0 = i, u_0 = u\big) &= p_0^{iu,t}(\bar{j} \mid i,u) \cdot \nu_1^{iu,t}(\bar{v} \mid \bar{j}), \qquad \text{for } k = 1,\\
\mathbf{P}_t^{iu}\big(i_k = \bar{j}, u_k = \bar{v} \mid i_{k-1} = \bar{i}, u_{k-1} = \bar{u}\big) &= p_{k-1}^{iu,t}(\bar{j} \mid \bar{i},\bar{u}) \cdot \nu_k^{iu,t}(\bar{v} \mid \bar{j}), \qquad \text{for } k \geq 2,
\end{aligned}$$

where for all $k \geq 1$ and $(\bar{i},\bar{u}) \in R_o$, $\bar{j} \in S_o$,

$$p_{k-1}^{iu,t}(\cdot \mid \bar{i},\bar{u}) \in A_\delta(\bar{i},\bar{u}), \qquad \nu_k^{iu,t}(\cdot \mid \bar{j}) \in \mathcal{P}(\mathcal{U}) \text{ with } \mathrm{supp}\left(\nu_k^{iu,t}(\cdot \mid \bar{j})\right) \subset U(\bar{j}),$$

and moreover, $p_0^{iu,t}(\cdot \mid i,u) = \mathbf{q}_t^{iu}$ if $t \geq t_0$.

(c) *The one-stage cost functions $g_k^{iu,t}$ in (a) satisfy*

$$g_k^{iu,t} \in B_\delta, \qquad k \geq 0,$$

and moreover, $g_0^{iu,t}(i,u) = \tilde{g}_t(i,u)$ if $t \geq t_0$.

(d) *For the Markov chain in (a), there exists an integer k_t such that $\{(i_k,u_k), k \geq k_t\}$ evolves and incurs costs as in the original SSP problem under the proper policy ν; i.e., for $k \geq k_t$,*

$$\nu_k^{iu,t}(\cdot \mid \bar{i}) = \nu(\cdot \mid \bar{i}), \qquad p_k^{iu,t}(\cdot \mid \bar{i},\bar{u}) = \mathbf{p}_o^{\bar{i}\bar{u}}, \qquad g_k^{iu,t}(\bar{i},\bar{u}) = g(\bar{i},\bar{u}), \quad \forall\, (\bar{i},\bar{u}) \in R_o.$$

Proof. The proof is by induction on t. For $t = t_0$, $\tilde{Q}_{t_0}$ satisfies properties (a)-(d) by its definition and our choice of the sample path and t_0 (cf. Lemma 3.2). [In particular, for each $(i,u) \in R$, p_k^{iu,t_0} and ν_k^{iu,t_0} in (a) are given by: for $k = 0$,

$$p_0^{iu,t_0}(\cdot \mid i,u) = \mathbf{q}_{t_0}^{iu}, \qquad p_0^{iu,t_0}(\cdot \mid \bar{i},\bar{u}) = \mathbf{p}_o^{\bar{i}\bar{u}}, \quad \forall\, (\bar{i},\bar{u}) \in R_o \setminus \{(i,u)\},$$

and for all $k \geq 1$,

$$p_k^{iu,t_0}(\cdot \mid \bar{i},\bar{u}) = \mathbf{p}_o^{\bar{i}\bar{u}}, \quad \forall\, (\bar{i},\bar{u}) \in R_o, \qquad \nu_k^{iu,t_0} = \nu,$$

whereas $k_{t_0} = 1$ in (d).] For $t < t_0$, since $\tilde{Q}_t = \tilde{Q}_{t_0}$ by definition, they also satisfy (a)-(d). So let us assume that properties (a)-(d) are satisfied by all $\tilde{Q}_\tau$, $0 \le \tau \le t$, for some $t \ge t_0$. We will show that $\tilde{Q}_{t+1}$ also has these properties.

Consider $\tilde{Q}_{t+1}(i,u)$ for each $(i,u) \in R$. To simplify notation, denote $\gamma = \gamma_t(i,u) \in [0,1]$ (cf. Lemma 3.2). By Eq. (3.6),

$$\tilde{Q}_{t+1}(i,u) = (1-\gamma)\,\tilde{Q}_t(i,u) + \gamma\,\big(g(i,u) + \omega_t(i,u) + \tilde{Q}_{\tau_t^{sv}(i,u)}(s,v)\big),$$

where $s = j_t^{iu}, v = v_t^{iu}$, and $\tau_t^{sv}(i,u) \le t$. By the induction hypothesis, $\tilde{Q}_t$ and $\tilde{Q}_{\tau_t^{sv}(i,u)}$ can be expressed as in (a), so denoting $\bar{\tau} = \tau_t^{sv}(i,u)$ for short and noticing $\mathbf{P}_t^{iu}(i_0 = i, u_0 = u) = 1$ by property (b), we have

$$\begin{aligned}
\tilde{Q}_{t+1}(i,u) &= (1-\gamma)\sum_{k=0}^{\infty} \mathrm{E}^{\mathbf{P}_t^{iu}}\Big[\, g_k^{iu,t}(i_k,u_k)\,\Big] + \gamma\,\big(g(i,u)+\omega_t(i,u)\big) + \gamma \sum_{k=0}^{\infty} \mathrm{E}^{\mathbf{P}_{\bar{\tau}}^{sv}}\Big[\, g_k^{sv,\bar{\tau}}(i_k,u_k)\,\Big] \\
&= (1-\gamma)\, g_0^{iu,t}(i,u) + \gamma\,\big(g(i,u)+\omega_t(i,u)\big) \\
&\quad + \sum_{k=1}^{\infty} \left\{ (1-\gamma)\,\mathrm{E}^{\mathbf{P}_t^{iu}}\Big[\, g_k^{iu,t}(i_k,u_k)\,\Big] + \gamma\,\mathrm{E}^{\mathbf{P}_{\bar{\tau}}^{sv}}\Big[\, g_{k-1}^{sv,\bar{\tau}}(i_{k-1},u_{k-1})\,\Big] \right\}, \\
&= \sum_{k\ge 0} C_k, \qquad (3.16)
\end{aligned}$$

where

$$C_0 = (1-\gamma)\, g_0^{iu,t}(i,u) + \gamma\,\big(g(i,u)+\omega_t(i,u)\big), \qquad (3.17)$$

$$C_k = (1-\gamma)\,\mathrm{E}^{\mathbf{P}_t^{iu}}\Big[\, g_k^{iu,t}(i_k,u_k)\,\Big] + \gamma\,\mathrm{E}^{\mathbf{P}_{\bar{\tau}}^{sv}}\Big[\, g_{k-1}^{sv,\bar{\tau}}(i_{k-1},u_{k-1})\,\Big], \qquad k \ge 1. \qquad (3.18)$$

Next we will rewrite each term C_k in a desirable form. During this procedure, we will construct the transition probabilities $p_k^{iu,t+1}$ and $\nu_k^{iu,t+1}$ that compose the probability distribution $\mathbf{P}_{t+1}^{iu}$ of the time-inhomogeneous Markov chain for $t+1$, as well as the one-stage cost functions $g_k^{iu,t+1}$ required in the lemma. For clarity we divide the rest of the proof in five steps.

(1) We consider the term C_0 in Eq. (3.17) and define the transition probabilities and one-stage costs for $k = 0$ and $t+1$. By the induction hypothesis and property (c), $g_0^{iu,t}(i,u) = \tilde{g}_t(i,u)$. Using this and the definition of $\{\tilde{g}_t\}$ [cf. Eq. (3.9)], we have

$$C_0 = (1-\gamma)\,\tilde{g}_t(i,u) + \gamma\,\big(g(i,u)+\omega_t(i,u)\big) = \tilde{g}_{t+1}(i,u). \qquad (3.19)$$

Let us define the cost function and transition probabilities for $k = 0$ and $t+1$ by

$$g_0^{iu,t+1} = \tilde{g}_{t+1}, \qquad p_0^{iu,t+1}(\cdot \mid i,u) = \mathbf{q}_{t+1}^{iu},$$

and

$$p_0^{iu,t+1}(\cdot \mid \bar{i},\bar{u}) = \mathbf{p}_o^{\bar{i}\bar{u}}, \qquad \forall\,(\bar{i},\bar{u}) \in R_o \setminus \{(i,u)\}.$$

By Lemma 3.2 and our choice of the sample path, $\tilde{g}_{t+1} \in B_\delta$ and $\mathbf{q}_{t+1}^{iu} \in A_\delta(i,u)$, so $g_0^{iu,t+1}$ and $p_0^{iu,t+1}$ satisfy the requirements in properties (b) and (c).

(2) We now consider the term C_k in Eq. (3.18), and we introduce several relations that will define the transition probabilities and one-stage costs for $k \ge 1$ and $t+1$ (the precise definitions will be given in the next two steps).

Consider each $k \geq 1$. Let P_1^k denote the law of (i_k, u_k, i_{k+1}) under $\mathbf{P}_t^{iu}$, and let P_2^k denote the law of (i_{k-1}, u_{k-1}, i_k) under $\mathbf{P}_{\bar{\tau}}^{sv}$. Let P_3^k denote the convex combination of them:

$$P_3^k = (1-\gamma)P_1^k + \gamma P_2^k.$$

We regard P_1^k, P_2^k, P_3^k as probability measures on the sample space $\tilde{\Omega} = S_o \times \mathcal{U} \times S_o$, and we denote by X, Y and Z the function that maps a point $(\bar{i}, \bar{u}, \bar{j}) \in \tilde{\Omega}$ to its 1st, 2nd and 3rd coordinate, respectively. By property (b) of $\mathbf{P}_t^{iu}$ and $\mathbf{P}_{\bar{\tau}}^{sv}$ from the induction hypothesis, it is clear that under either P_1^k or P_2^k, the possible values of (X, Y) are from the set R_o of state and feasible control pairs, so the subset $R_o \times S_o$ of $\tilde{\Omega}$ has probability 1 under P_3^k. Thus we can write C_k in Eq. (3.18) equivalently as

$$C_k = \sum_{\bar{i} \in S_o} \sum_{\bar{u} \in U(\bar{i})} \Big((1-\gamma)\, P_1^k(X = \bar{i}, Y = \bar{u}) \cdot g_k^{iu,t}(\bar{i}, \bar{u}) + \gamma\, P_2^k(X = \bar{i}, Y = \bar{u}) \cdot g_{k-1}^{sv,\bar{\tau}}(\bar{i}, \bar{u}) \Big). \quad (3.20)$$

In the next two steps, we will introduce one-stage cost functions $g_k^{iu,t+1}$ to rewrite Eq. (3.20) equivalently as

$$C_k = \sum_{\bar{i} \in S_o} \sum_{\bar{u} \in U(\bar{i})} P_3^k(X = \bar{i}, Y = \bar{u}) \cdot g_k^{iu,t+1}(\bar{i}, \bar{u}). \quad (3.21)$$

We will also define the transition probabilities $\nu_k^{iu,t+1}(\cdot \mid \bar{i})$ and $p_k^{iu,t+1}(\cdot \mid \bar{i}, \bar{u})$ to express P_3^k as

$$P_3^k(X = \bar{i}, Y = \bar{u}) = P_3^k(X = \bar{i}) \cdot \nu_k^{iu,t+1}(\bar{u} \mid \bar{i}), \quad (3.22)$$

$$P_3^k(X = \bar{i}, Y = \bar{u}, Z = \bar{j}) = P_3^k(X = \bar{i}, Y = \bar{u}) \cdot p_k^{iu,t+1}(\bar{j} \mid \bar{i}, \bar{u}), \quad (3.23)$$

for all $(\bar{i}, \bar{u}) \in R_o$ and $\bar{j} \in S_o$. Note that in the above, by the definition of P_3^k,

$$P_3^k(X = \bar{i}) = (1-\gamma)\, \mathbf{P}_t^{iu}\big(i_k = \bar{i}\,\big) + \gamma\, \mathbf{P}_{\bar{\tau}}^{sv}\big(i_{k-1} = \bar{i}\,\big), \qquad \forall\ \bar{i} \in S_o. \quad (3.24)$$

(3) We now define the one-stage cost functions for $k \geq 1$ and $t+1$.

Consider each $k \geq 1$. Define the cost function $g_k^{iu,t+1}$ as follows: for each $(\bar{i}, \bar{u}) \in R_o$,

$$g_k^{iu,t+1}(\bar{i}, \bar{u}) = \frac{(1-\gamma)\, P_1^k(X = \bar{i}, Y = \bar{u})}{P_3^k(X = \bar{i}, Y = \bar{u})} \cdot g_k^{iu,t}(\bar{i}, \bar{u}) + \frac{\gamma\, P_2^k(X = \bar{i}, Y = \bar{u})}{P_3^k(X = \bar{i}, Y = \bar{u})} \cdot g_{k-1}^{sv,\bar{\tau}}(\bar{i}, \bar{u}) \quad (3.25)$$

if $P_3^k(X = \bar{i}, Y = \bar{u}) > 0$, and $g_k^{iu,t+1}(\bar{i}, \bar{u}) = g(\bar{i}, \bar{u})$ otherwise. With this definition, the expression for C_k given in Eq. (3.21) is clearly true and equivalent to that given in Eq. (3.20).

We verify that $g_k^{iu,t+1}$ satisfies the requirement in property (c), that is,

$$g_k^{iu,t+1} \in B_\delta. \quad (3.26)$$

Consider each $(\bar{i}, \bar{u}) \in R_o$ and discuss two cases. If $P_3^k(X = \bar{i}, Y = \bar{u}) = 0$, then $|g_k^{iu,t+1}(\bar{i}, \bar{u}) - g(\bar{i}, \bar{u})| = 0$ by definition. Suppose $P_3^k(X = \bar{i}, Y = \bar{u}) > 0$. Then by Eq. (3.25), $g_k^{iu,t+1}(\bar{i}, \bar{u})$ is a convex combination of $g_k^{iu,t}(\bar{i}, \bar{u})$ and $g_{k-1}^{sv,\bar{\tau}}(\bar{i}, \bar{u})$, whereas $g_k^{iu,t}, g_{k-1}^{sv,\bar{\tau}} \in B_\delta$ by the induction hypothesis (property (c)). This implies, by the definition of B_δ, that $|g_k^{iu,t+1}(\bar{i}, \bar{u}) - g(\bar{i}, \bar{u})| \leq \delta$ for $(\bar{i}, \bar{u}) \in R$ and $g_k^{iu,t+1}(0, 0) = 0$ for $(\bar{i}, \bar{u}) = (0, 0)$. Combining the two cases, and in view also of the definition of B_δ, we have that $g_k^{iu,t+1}$ satisfies Eq. (3.26).

We verify that $g_k^{iu,t+1}$ satisfies the requirement in property (d). By the induction hypothesis $g_k^{iu,t} = g$ for $k \geq k_t$ and $g_{k-1}^{sv,\bar{\tau}} = g$ for $k \geq k_{\bar{\tau}} + 1$, whereas each component of $g_k^{iu,t+1}$ by definition

either equals the corresponding component of g or is a convex combination of the corresponding components of $g_k^{iu,t}$ and $g_{k-1}^{sv,\bar{\tau}}$. Hence

$$g_k^{iu,t+1} = g, \qquad \forall\, k \geq k_{t+1} \stackrel{\text{def}}{=} \max\{k_t, k_{\bar{\tau}} + 1\}. \tag{3.27}$$

(4) We now define the transition probabilities for $k \geq 1$ and $t+1$.

Consider each $k \geq 1$. Define the transition probability distributions $\nu_k^{iu,t+1}$ and $p_k^{iu,t+1}$ as follows:

$$\nu_k^{iu,t+1}(\cdot \mid \bar{i}) = P_3^k(Y = \cdot \mid X = \bar{i}), \qquad \forall\, \bar{i} \in S_o, \tag{3.28}$$

$$p_k^{iu,t+1}(\cdot \mid \bar{i}, \bar{u}) = P_3^k(Z = \cdot \mid X = \bar{i}, Y = \bar{u}), \qquad \forall\, (\bar{i}, \bar{u}) \in R_o. \tag{3.29}$$

If in the right-hand sides of Eqs. (3.28)-(3.29), an event being conditioned upon has probability zero, then let the corresponding conditional probability (which can be defined arbitrarily) be defined according to the following:

$$\begin{aligned} P_3^k(Y = \cdot \mid X = \bar{i}) &= \nu(\cdot \mid \bar{i}), \qquad \text{if } \; P_3^k(X = \bar{i}) = 0; \\ P_3^k(Z = \cdot \mid X = \bar{i}, Y = \bar{u}) &= \mathbf{p}_o^{\bar{i}\bar{u}}, \qquad \text{if } \; P_3^k(X = \bar{i}, Y = \bar{u}) = 0. \end{aligned}$$

With the above definitions, the equalities (3.22) and (3.23) desired in step (2) of the proof clearly hold. We now verify that $\nu_k^{iu,t+1}$ and $p_k^{iu,t+1}$ satisfy the requirements in properties (b) and (d).

First, we show that $p_k^{iu,t+1}$ satisfies the requirement in property (b), that is,

$$p_k^{iu,t+1}(\cdot \mid \bar{i}, \bar{u}) \in A_\delta(\bar{i}, \bar{u}), \qquad \forall\, (\bar{i}, \bar{u}) \in R_o.$$

This holds by the definition of $p_k^{iu,t+1}(\cdot \mid \bar{i}, \bar{u})$ if $P_3^k(X = \bar{i}, Y = \bar{u}) = 0$, so let us consider the case $P_3^k(X = \bar{i}, Y = \bar{u}) > 0$ for each $(\bar{i}, \bar{u}) \in R_o$. By the induction hypothesis $\mathbf{P}_t^{iu}$ and $\mathbf{P}_{\bar{\tau}}^{sv}$ satisfy property (b). Using this and the definition of P_1^k and P_2^k, we have that for all $\bar{j} \in S_o$,

$$\begin{aligned} P_1^k(X = \bar{i}, Y = \bar{u}, Z = \bar{j}) &= \mathbf{P}_t^{iu}\big(i_k = \bar{i}, u_k = \bar{u}\big) \cdot p_k^{iu,t}(\bar{j} \mid \bar{i}, \bar{u}), \\ P_2^k(X = \bar{i}, Y = \bar{u}, Z = \bar{j}) &= \mathbf{P}_{\bar{\tau}}^{sv}\big(i_{k-1} = \bar{i}, u_{k-1} = \bar{u}\big) \cdot p_{k-1}^{sv,\bar{\tau}}(\bar{j} \mid \bar{i}, \bar{u}), \end{aligned}$$

which implies

$$P_1^k(Z = \cdot \mid X = \bar{i}, Y = \bar{u}) = p_k^{iu,t}(\cdot \mid \bar{i}, \bar{u}), \qquad P_2^k(Z = \cdot \mid X = \bar{i}, Y = \bar{u}) = p_{k-1}^{sv,\bar{\tau}}(\cdot \mid \bar{i}, \bar{u}), \tag{3.30}$$

and by property (b) from the induction hypothesis again,

$$P_1^k(Z = \cdot \mid X = \bar{i}, Y = \bar{u}) \in A_\delta(\bar{i}, \bar{u}), \qquad P_2^k(Z = \cdot \mid X = \bar{i}, Y = \bar{u}) \in A_\delta(\bar{i}, \bar{u}). \tag{3.31}$$

Then, since $P_3^k = (1-\gamma)P_1^k + \gamma P_2^k$ with $\gamma \in [0,1]$, we have

$$\begin{aligned} P_3^k(Z = \cdot \mid X = \bar{i}, Y = \bar{u}) &= \frac{P_3^k(X = \bar{i}, Y = \bar{u}, Z = \cdot)}{P_3^k(X = \bar{i}, Y = \bar{u})} \\ &= \big(1 - \beta(\bar{i}, \bar{u})\big) \cdot P_1^k(Z = \cdot \mid X = \bar{i}, Y = \bar{u}) + \beta(\bar{i}, \bar{u}) \cdot P_2^k(Z = \cdot \mid X = \bar{i}, Y = \bar{u}), \end{aligned} \tag{3.32}$$

where

$$\beta(\bar{i}, \bar{u}) = \frac{\gamma P_2^k(X = \bar{i}, Y = \bar{u})}{(1-\gamma)P_1^k(X = \bar{i}, Y = \bar{u}) + \gamma P_2^k(X = \bar{i}, Y = \bar{u})} \; \in [0,1].$$

Since the set $A_\delta(\bar{i}, \bar{u})$ is convex, using the fact $\beta(\bar{i}, \bar{u}) \in [0,1]$, Eqs. (3.31)-(3.32) imply that

$$P_3^k(Z = \cdot \mid X = \bar{i}, Y = \bar{u}) \in A_\delta(\bar{i}, \bar{u}),$$

and therefore, by definition [cf. Eq. (3.29)], $p_k^{iu,t+1}(\cdot \mid \bar{i}, \bar{u}) = P_3^k(Z = \cdot \mid X = \bar{i}, Y = \bar{u}) \in A_\delta(\bar{i}, \bar{u})$.

We now verify that $p_k^{iu,t+1}$ satisfies the requirement in property (d): for all $(\bar{i}, \bar{u}) \in R_o$,

$$p_k^{iu,t+1}(\cdot \mid \bar{i}, \bar{u}) = \mathbf{p}_o^{\bar{i}\bar{u}}, \qquad \forall\, k \geq k_{t+1} = \max\{k_t, k_{\bar{\tau}} + 1\}. \tag{3.33}$$

By the induction hypothesis, property (d) is satisfied for $\tau \leq t$, and in particular, for all $(\bar{i}, \bar{u}) \in R_o$, $p_k^{iu,t}(\cdot \mid \bar{i}, \bar{u}) = \mathbf{p}_o^{\bar{i}\bar{u}}$ for $k \geq k_t$ and $p_k^{sv,\bar{\tau}}(\cdot \mid \bar{i}, \bar{u}) = \mathbf{p}_o^{\bar{i}\bar{u}}$ for $k \geq k_{\bar{\tau}}$. In view of Eqs. (3.30) and (3.32), we have that if $P_3^k(X = \bar{i}, Y = \bar{u}) > 0$, then $p_k^{iu,t+1}(\cdot \mid \bar{i}, \bar{u})$ is a convex combination of $p_k^{iu,t}(\cdot \mid \bar{i}, \bar{u})$ and $p_{k-1}^{sv,\bar{\tau}}(\cdot \mid \bar{i}, \bar{u})$ and hence satisfies Eq. (3.33). But if $P_3^k(X = \bar{i}, Y = \bar{u}) = 0$, $p_k^{iu,t+1}(\cdot \mid \bar{i}, \bar{u}) = \mathbf{p}_o^{\bar{i}\bar{u}}$ by definition. Hence Eq. (3.33) holds.

We now verify that $\nu_k^{iu,t+1}$ given by Eq. (3.28) satisfies the requirements in properties (b) and (d). For each $\bar{i} \in S_o$, $\nu_k^{iu,t+1}(\cdot \mid \bar{i}) = \nu(\cdot \mid \bar{i})$ by definition if $P_3^k(X = \bar{i}) = 0$; otherwise, similar to the preceding proof, $\nu_k^{iu,t+1}(\cdot \mid \bar{i})$ can be expressed as a convex combination of $\nu_k^{iu,t}(\cdot \mid \bar{i})$ and $\nu_{k-1}^{sv,\bar{\tau}}(\cdot \mid \bar{i})$:

$$\nu_k^{iu,t+1}(\cdot \mid \bar{i}) = \frac{(1-\gamma)P_1^k(X = \bar{i})}{P_3^k(X = \bar{i})} \cdot \nu_k^{iu,t}(\cdot \mid \bar{i}) + \frac{\gamma P_2^k(X = \bar{i})}{P_3^k(X = \bar{i})} \cdot \nu_{k-1}^{sv,\bar{\tau}}(\cdot \mid \bar{i}),$$

where if $k = 1$ and $\bar{i} = s$, we let $\nu_0^{sv,\bar{\tau}}(\cdot \mid s)$ denote the distribution in $\mathcal{P}(\mathcal{U})$ that assigns probability 1 to the control v [if $k = 1$ and $\bar{i} \neq s$, then the second term above is zero because $\mathbf{P}_{\bar{\tau}}^{sv}(i_0 = s, u_0 = v) = 1$ by the induction hypothesis and consequently, $P_2^1(X = \bar{i}) = \mathbf{P}_{\bar{\tau}}^{sv}(i_0 = \bar{i}) = 0$]. Combining the two cases, and using properties (b) and (d) of the induction hypothesis, we then have that $\operatorname{supp}\big(\nu_k^{iu,t+1}(\cdot \mid \bar{i})\big) \subset U(\bar{i})$ for $\bar{i} \in S_o$, and

$$\nu_k^{iu,t+1}(\cdot \mid \bar{i}) = \nu(\cdot \mid \bar{i}), \qquad \forall\, k \geq k_{t+1},\ \bar{i} \in S_o, \tag{3.34}$$

which are the requirements for $\nu_k^{iu,t+1}$ in properties (b) and (d).

(5) In this last step of the proof, we define the Markov chain for $t+1$ and verify the expression for $\tilde{Q}_{t+1}(i, u)$ given in property (a).

Let the time-inhomogeneous Markov chain $\{(i_k, u_k), k \geq 0\}$ with probability distribution $\mathbf{P}_{t+1}^{iu}$, required in property (a) for $t+1$, be as follows. Let the chain start with $(i_0, u_0) = (i, u)$, and let its transition probabilities have the product forms given in property (b) for $t+1$, where $p_k^{iu,t+1}, k \geq 0$, and $\nu_k^{iu,t+1}, k \geq 1$, are the functions that we defined in the preceding proof. Also let the time-varying one-stage cost functions $g_k^{iu,t+1}, k \geq 0$, be as defined earlier. We have shown that these transition probabilities and one-stage cost functions satisfy the requirements in properties (b)-(d). To prove the lemma, what we still need to show is that with our definitions, the expression given in property (a) equals $\tilde{Q}_{t+1}(i, u)$.

First of all, because our definitions of the transition probabilities and one-stage cost functions for $t+1$ satisfy property (d), they ensure that under $\mathbf{P}_{t+1}^{iu}$, $\{(i_k, u_k), k \geq k_{t+1}\}$ evolves and incurs costs as in the original SSP problem under the proper stationary policy ν. Consequently, $\mathrm{E}^{\mathbf{P}_{t+1}^{iu}}\Big[\sum_{k=0}^{\infty} g_k^{iu,t+1}(i_k, u_k)\Big]$ is well-defined and finite, and the order of summation and expectation can be exchanged (the reason is the same as the one we gave at the end of Section 3.3.2 for the expression of $\tilde{Q}_{t_0}$):

$$\mathrm{E}^{\mathbf{P}_{t+1}^{iu}}\Big[\sum_{k=0}^{\infty} g_k^{iu,t+1}(i_k, u_k)\Big] = \sum_{k=0}^{\infty} \mathrm{E}^{\mathbf{P}_{t+1}^{iu}}\Big[g_k^{iu,t+1}(i_k, u_k)\Big]. \tag{3.35}$$

Hence, to prove property (a) for $t+1$, that is, to show

$$\tilde{Q}_{t+1}(i, u) = g_0^{iu,t+1}(i, u) + \sum_{k=1}^{\infty} \mathrm{E}^{\mathbf{P}_{t+1}^{iu}}\Big[g_k^{iu,t+1}(i_k, u_k)\Big],$$

we only need to show, in view of the fact $\tilde{Q}_{t+1}(i,u) = \sum_{k=0}^{\infty} C_k$ [cf. Eq. (3.16)], that

$$C_0 = g_0^{iu,t+1}(i,u), \qquad C_k = \mathrm{E}^{\mathbf{P}_{t+1}^{iu}}\Big[g_k^{iu,t+1}(i_k,u_k) \Big], \quad k \geq 1. \tag{3.36}$$

The first relation is true since by definition $g_0^{iu,t+1}(i,u) = \tilde{g}_{t+1}(i,u) = C_0$ [cf. Eq. (3.19)]. We now prove the second equality for $C_k, k \geq 1$.

For $k \geq 1$, recall that by Eq. (3.21),

$$C_k = \sum_{\bar{i} \in S_o} \sum_{\bar{u} \in U(\bar{i})} P_3^k(X = \bar{i}, Y = \bar{u}) \cdot g_k^{iu,t+1}(\bar{i},\bar{u}).$$

Hence, to prove the desired equality for C_k, it is sufficient to prove that

$$\mathbf{P}_{t+1}^{iu}(i_k = \bar{i}, u_k = \bar{u}) = P_3^k(X = \bar{i}, Y = \bar{u}), \qquad \forall\, (\bar{i},\bar{u}) \in R_o. \tag{3.37}$$

By the definition of $\mathbf{P}_{t+1}^{iu}$, $\mathbf{P}_{t+1}^{iu}(u_k = \bar{u} \mid i_k = \bar{i}) = \nu_k^{iu,t+1}(\bar{u} \mid \bar{i})$ for all $(\bar{i},\bar{u}) \in R_o$, so, in view of Eq. (3.22), the equality (3.37) will be implied if we prove

$$\mathbf{P}_{t+1}^{iu}(i_k = \bar{i}) = P_3^k(X = \bar{i}), \qquad \forall\, \bar{i} \in S_o. \tag{3.38}$$

We verify Eq. (3.38) by induction on k. For $k = 1$, using Eq. (3.24) and property (b) of $\mathbf{P}_t^{iu}$ and $\mathbf{P}_{\bar{\tau}}^{sv}$, we have that for every $\bar{i} \in S_o$,

$$\begin{aligned} P_3^1(X = \bar{i}) &= (1-\gamma)\, \mathbf{P}_t^{iu}\big(i_1 = \bar{i}\big) + \gamma\, \mathbf{P}_{\bar{\tau}}^{sv}\big(i_0 = \bar{i}\big) \\ &= (1-\gamma)\, p_0^{iu,t}(\bar{i} \mid i,u) + \gamma\, \mathbf{e}_s(\bar{i}) \\ &= (1-\gamma)\, \mathbf{q}_t^{iu}(\bar{i}) + \gamma\, \mathbf{e}_{j_t^{iu}}(\bar{i}) \\ &= \mathbf{q}_{t+1}^{iu}(\bar{i}) = p_0^{iu,t+1}(\bar{i} \mid i,u) = \mathbf{P}_{t+1}^{iu}(i_1 = \bar{i}), \end{aligned}$$

where the last three equalities follow from the definition of $\mathbf{q}_{t+1}^{iu}$ [cf. Eq. (3.10)], the definition of $p_0^{iu,t+1}$, and the definition of $\mathbf{P}_{t+1}^{iu}$, respectively. Hence Eq. (3.38) holds for $k = 1$.

Suppose Eq. (3.38) holds for some $k \geq 1$. Then, by the definition of $\mathbf{P}_{t+1}^{iu}$, we have that for all $\bar{j} \in S_o$,

$$\begin{aligned} \mathbf{P}_{t+1}^{iu}(i_{k+1} = \bar{j}) &= \sum_{\bar{i} \in S_o} \sum_{\bar{u} \in U(\bar{i})} \mathbf{P}_{t+1}^{iu}(i_k = \bar{i}) \cdot \nu_k^{iu,t+1}(\bar{u} \mid \bar{i}) \cdot p_k^{iu,t+1}(\bar{j} \mid \bar{i},\bar{u}) \\ &= \sum_{\bar{i} \in S_o} \sum_{\bar{u} \in U(\bar{i})} P_3^k(X = \bar{i}) \cdot \nu_k^{iu,t+1}(\bar{u} \mid \bar{i}) \cdot p_k^{iu,t+1}(\bar{j} \mid \bar{i},\bar{u}) \\ &= P_3^k(Z = \bar{j}) = P_3^{k+1}(X = \bar{j}), \end{aligned}$$

where the second equality follows from the induction hypothesis, the third equality follows from Eqs. (3.22)-(3.23), and the last equality follows from the definition of P_3^k and P_3^{k+1}. This completes the induction and proves Eq. (3.38) for all $k \geq 1$, which in turn establishes Eq. (3.37) for all $k \geq 1$. Consequently, for all $k \geq 1$, the desired equality (3.36) for C_k holds, and we conclude that $\tilde{Q}_{t+1}(i,u)$ equals the expressions given in Eq. (3.35). This completes the proof of the lemma. □

3.3.4 Lower boundedness of $\{\tilde{Q}_t\}$

In Sections 3.3.2 and 3.3.3, we have shown that for each sample path from a set of probability one, and for each $\delta > 0$, we can construct a sequence $\{\tilde{Q}_t\}$ such that $\tilde{Q}_t(i,u)$ for each $(i,u) \in R$ is the expected total cost of a randomized Markov policy in an MDP that has time-varying transition and

one-stage cost parameters lying in the δ-neighborhood of the respective parameters of the original SSP problem. By Lemma 3.1, therefore, to complete the boundedness proof for the Q-learning iterates $\{Q_t\}$, it is sufficient to show that when δ is sufficiently small, the expected total costs of all policies in all these neighboring MDPs cannot be unbounded from below.

The latter can in turn be addressed by considering the following total cost MDP. It has the same state space S_o with state 0 being absorbing and cost-free. For each state $i \in S$, the set of feasible controls consists of not only the regular controls $U(i)$, but also the transition probabilities and one-stage cost functions. More precisely, the extended control set at state i is defined to be

$$U_\delta(i) = \big\{(u, \mathbf{p}^{iu}, \theta_i) \,\big|\, u \in U(i), \mathbf{p}^{iu} \in A_\delta(i,u), \theta_i \in B_\delta(i)\big\},$$

where $B_\delta(i)$ is a set of one-stage cost functions at i: with $z = \{z(u), u \in U(i)\}$,

$$B_\delta(i) = \big\{z \,\big|\, |z(u) - g(i,u)| \le \delta, \forall\, u \in U(i)\big\}.$$

Applying control $(u, \mathbf{p}^{iu}, \theta_i)$ at $i \in S$, the one-stage cost, denoted by $c(u; i, \theta_i)$, is

$$c(u; i, \theta_i) = \theta_i(u),$$

and the probability of transition from state i to j is $\mathbf{p}^{iu}(j)$. We refer to this problem as the *extended SSP problem*. If we can show that the optimal total costs of this problem for all initial states are finite, then it will imply that $\{\tilde{Q}_t\}$ is bounded below because by Lemma 3.3, for each t and $(i,u) \in R$, $\tilde{Q}_t(i,u)$ equals the expected total cost of some policy in the extended SSP problem for the initial state i.

The extended SSP problem has a finite number of states and a compact control set for each state. Its one-stage cost $c(u; i, \theta_i)$ is a continuous function of the control component (u, θ_i), whereas its transition probabilities are continuous functions of the control component $(u, \mathbf{p}^{iu})$ for each state i. With these compactness and continuity properties, the extended SSP problem falls into the set of SSP models analyzed in [BT91]. Based on the results of [BT91], the optimal total cost function of the extended SSP problem is finite everywhere if Assumption 1.1 holds in this problem – that is, if the extended SSP problem satisfies the following two conditions: (i) there exists at least one proper deterministic stationary policy, and (ii) any improper deterministic stationary policy incurs infinite cost for some initial state.

Lemma 3.4 ([BT91]). *If the extended SSP problem satisfies Assumption 1.1, then its optimal total cost is finite for every initial state.*

The extended SSP problem clearly has at least one proper deterministic stationary policy, which is to apply at a state $i \in S$ the control $\big(\mu(i), \mathbf{p}_o^{i\mu(i)}, g(i,\cdot)\big)$, where μ is a proper policy in the set $\Pi_{\rm SD}$ of the original SSP problem (such a policy exists in view of Assumption 1.1(i) on the original SSP problem). We now show that for sufficiently small δ, any improper deterministic stationary policy of the extended SSP problem incurs infinite cost for some initial state.

To this end, let us restrict δ to be no greater than some $\delta_0 > 0$, for which $p_{ij}(u) > 0$ implies $\mathbf{p}^{iu}(j) > 0$ for all $\mathbf{p}^{iu} \in A_\delta(i,u)$ and $(i,u) \in R$, i.e.,

$$\mathbf{p}_o^{iu} \ll \mathbf{p}^{iu}, \qquad \forall\, \mathbf{p}^{iu} \in A_\delta(i,u),\ (i,u) \in R,\ \delta \le \delta_0. \tag{3.39}$$

[Recall that we also have $\mathbf{p}^{iu} \ll \mathbf{p}_o^{iu}$ in view of the definition of $A_\delta(i,u)$.] To simplify notation, denote

$$\mathcal{A}_\delta = \bigtimes_{(i,u)\in R} A_\delta(i,u).$$

Recall the definition of the set B_δ, which is a subset of vectors in the δ-neighborhood of the expected one-stage cost vector g of the original problem: with $\mathbf{c} = \{c(i,u),\, (i,u) \in R_o\}$,

$$B_\delta = \big\{\, \mathbf{c} \,\big|\, c(0,0) = 0 \text{ and } |c(i,u) - g(i,u)| \le \delta,\ \forall\, (i,u) \in R\big\}.$$

Note that $B_\delta = \bigtimes_{i \in S_o} B_\delta(i)$, where $B_\delta(0) = \{0\}$ and $B_\delta(i), i \in S$ are as defined earlier [for the control sets $U_\delta(i)$ of the extended SSP problem]. For each $\Gamma \in \mathcal{A}_\delta$ and $\theta \in B_\delta$, let us call an MDP *a perturbed SSP problem with parameters* (Γ, θ), if it is the same as the original SSP problem except that the transition probabilities and one-stage costs for $(i, u) \in R$ are given by the respective components of Γ and θ.

Consider now a deterministic and stationary policy ζ of the extended SSP problem, which applies at each state i some feasible control $\zeta(i) = (\mu(i), \mathbf{p}^{i\mu(i)}, \theta_i) \in U_\delta(i)$. The regular controls $\mu(i)$ that ζ applies at states i correspond to a deterministic stationary policy of the original SSP problem, which we denote by μ. Then, by Eq. (3.39), ζ is proper (or improper) in the extended SSP problem if and only if μ is proper (or improper) in the original SSP problem. This is because by Eq. (3.39), the topology of the transition graph of the Markov chain on S_o that ζ induces in the extended SSP problem is the same as that of the Markov chain induced by μ in the original SSP problem, regardless of the two other control components $(\mathbf{p}^{i\mu(i)}, \theta_i)$ of ζ. Therefore, for Assumption 1.1(ii) to hold in the extended SSP problem, it is sufficient that any improper policy μ in Π_{SD} of the original problem has infinite cost for at least one initial state, in all perturbed SSP problems with parameters $\Gamma \in \mathcal{A}_\delta$ and $\theta \in B_\delta$ [cf. the relation between $\mathcal{A}_\delta, B_\delta$ and the control sets $U_\delta(i)$]. The next lemma shows that the latter is true for sufficiently small δ, thus providing the result we want.

Lemma 3.5. *Suppose the original SSP problem satisfies Assumption 1.1(ii). Then there exists $\delta_1 \in (0, \delta_0]$, where δ_0 is as given in Eq. (3.39), such that for all $\delta \le \delta_1$, the following holds: for any improper policy $\mu \in \Pi_{\mathrm{SD}}$ of the original problem, there exists a state i (depending on μ) with*

$$\liminf_{k \to \infty} \tilde{J}_k^\mu(i; \Gamma, \theta) = +\infty, \qquad \forall\, \Gamma \in \mathcal{A}_\delta,\ \theta \in B_\delta,$$

where $\tilde{J}_k^\mu(\cdot; \Gamma, \theta)$ is the k-stage cost function of μ in the perturbed SSP problem with parameters (Γ, θ).

For the proof, we will use a relation between the long-run average cost of a stationary policy and the total cost of that policy, and we will also use a continuity property of the average cost with respect to perturbations of transition probabilities and one-stage costs. The next two lemmas state two facts that will be used in our proof.

Lemma 3.6. *Let μ be a stationary policy of a finite-space MDP. If the average cost of μ is non-positive (strictly positive, respectively) for an initial state, then its total cost is less than (equal to, respectively) $+\infty$ for that initial state.*

Proof. This follows from the inequalities of [Put94, Theorem 9.4.1(a), p. 472] applied to a single policy μ. □

For an irreducible finite-state Markov chain with transition probability matrix P, we say P is irreducible, and we let $\sigma(P)$ denote the unique invariant distribution, viewed as a vector.

Lemma 3.7. *For any irreducible transition matrix $\bar{P}$, $\sigma(\cdot)$ is a continuous function on a neighborhood of $\bar{P}$ in the space of transition matrices.*

Proof. Since $\bar{P}$ is irreducible, there exists a neighborhood $\mathcal{N}(\bar{P})$ in the space of transition matrices such that all $P \in \mathcal{N}(\bar{P})$ are irreducible. Fix some $\beta \in (0, 1)$ and denote $P_\beta = (1 - \beta)P + \beta I$ for a transition matrix P. Then, for all $P \in \mathcal{N}(\bar{P})$, P_β is irreducible and aperiodic and $\sigma(P) = \sigma(P_\beta)$ with strictly positive components; furthermore, by [Sen81, Proof of Theorem 1.1(f), p. 5-6], each row of the adjoint matrix $\mathrm{Adj}(I - P_\beta)$ is a left eigenvector of P_β corresponding to the eigenvalue 1, so any row of $\mathrm{Adj}(I - P_\beta)$ normalized by the sum of that row is $\sigma(P_\beta) = \sigma(P)$. Since $\mathrm{Adj}(I - P_\beta)$ is a continuous function of P, this shows that $\sigma(P)$ is continuous on $\mathcal{N}(\bar{P})$. □

Proof of Lemma 3.5. Since the set Π_{SD} of the original SSP problem is finite, the number of improper policies in this set is also finite. Therefore, it is sufficient to consider each improper policy in Π_{SD} and show that the claim holds for all δ no greater than some $\bar{\delta} > 0$.

Let $\delta \leq \delta_0$ and let $\mu \in \Pi_{\text{SD}}$ be an improper policy. In any perturbed problem with parameters $(\Gamma, \theta) \in \mathcal{A}_\delta \times B_\delta$, the topology of the transition graph of the Markov chain on S_o induced by μ is the same as that in the original problem. This means that the recurrent classes of the Markov chains induced by μ are also the same for all these δ-perturbed problems and the original problem. Since μ is an improper policy, it induces more than one recurrent classes, and on at least one of them, which we denote by E, the long-run average cost of μ in the original problem is strictly positive. The latter follows from Lemma 3.6 and the assumption that any improper policy incurs infinite cost for some initial state in the original problem (Assumption 1.1(ii)). Let us show that for δ sufficiently small, the average cost of μ on the recurrent class E in any perturbed problem with parameters $(\Gamma, \theta) \in \mathcal{A}_\delta \times B_\delta$ must also be strictly positive.

To this end, let $\bar{P}$ denote the transition matrix of the Markov chain on E induced by μ in the original problem. Let $\bar{\theta} = g$, the parameter of one-stage costs that corresponds to the original problem. For any one-stage costs parameter $\theta = \{\theta(i,u), (i,u) \in R_o\}$, let $c_E(\theta)$ denote the vector of one-stage costs, $\{\theta\big(i, \mu(i)\big), i \in E\}$, for states in E. Note that $c_E(\theta)$ is a continuous function of θ.

The matrix $\bar{P}$ is irreducible, so by Lemma 3.7, in the space of transition matrices on E, there exists a neighborhood $\mathcal{N}(\bar{P})$ of $\bar{P}$ on which $\sigma(P)$ is a (well-defined) continuous function. Consequently, on $\mathcal{N}(\bar{P}) \times \mathcal{N}(\bar{\theta})$, where $\mathcal{N}(\bar{\theta})$ is some neighborhood of $\bar{\theta}$, the scalar product $\sigma(P)c_E(\theta)$ is a continuous function of (P, θ) (where $\sigma(P)$, $c_E(\theta)$ are treated as a row and column vector, respectively). We have $\sigma(\bar{P})c_E(\bar{\theta}) > 0$ because in the original problem, the average cost of μ for any initial state in E is strictly positive, as we showed earlier. Therefore, there exists some neighborhood $\mathcal{N}(\bar{P}, \bar{\theta})$ of $(\bar{P}, \bar{\theta})$ contained in $\mathcal{N}(\bar{P}) \times \mathcal{N}(\bar{\theta})$, on which $\sigma(P)c_E(\theta) > 0$.

Let P_Γ denote the transition matrix of the Markov chain on E induced by μ for the perturbed problem with parameters (Γ, θ). There exists $\bar{\delta} > 0$ sufficiently small such that for all $\Gamma \in \mathcal{A}_{\bar{\delta}}$ and $\theta \in B_{\bar{\delta}}$, $(P_\Gamma, \theta) \in \mathcal{N}(\bar{P}, \bar{\theta})$. Then, for any perturbed problem with parameters $\Gamma \in \mathcal{A}_{\bar{\delta}}$ and $\theta \in B_{\bar{\delta}}$, since the average cost of μ for any initial state $i \in E$ is $\sigma(P_\Gamma)c_E(\theta) > 0$, we have by Lemma 3.6 that $\liminf_{k\to\infty} \tilde{J}_k^\mu(i; \Gamma, \theta) = +\infty$ for all states $i \in E$. The proof is now complete. □

With Lemma 3.5, we have established that if the original SSP problem satisfies Assumption 1.1, then the extended SSP problem satisfies Assumption 1.1 and hence, by Lemma 3.4, has finite optimal total costs for all initial states. As we showed earlier, combined with Lemma 3.3, this implies the lower boundedness of $\{\tilde{Q}_t\}$ for sufficiently small δ, stated in the following lemma, and thus completes our proof of Prop. 3.3 on the lower boundedness of $\{Q_t\}$.

Lemma 3.8. *Let Assumptions 1.1 and 2.2 hold. Let $\delta \in (0, \delta_1]$ where δ_1 is as given in Lemma 3.5. Then, on any sample path from the set of probability one given in Lemma 3.2, with t_0 and $\tilde{Q}_0$ defined as in Section 3.3.2 for the chosen δ, the sequence $\{\tilde{Q}_t\}$ defined by Eqs. (3.5)-(3.7) is bounded below.*

Proof of Prop. 3.3. The proposition follows from Lemma 3.8 and Lemma 3.1. □

We have now established Theorem 3.1 on the boundedness of the Q-learning iterates $\{Q_t\}$.

4 An Illustrative Example

In this section we consider a simple 3-states example shown in the left graph of Fig. 1. We use it to illustrate the randomized Markov policies and the time-inhomogeneous SSP problems associated with the auxiliary sequence $\{\tilde{Q}_t\}$, which we constructed in Section 3.3 for proving the lower boundedness of the Q-learning iterates.

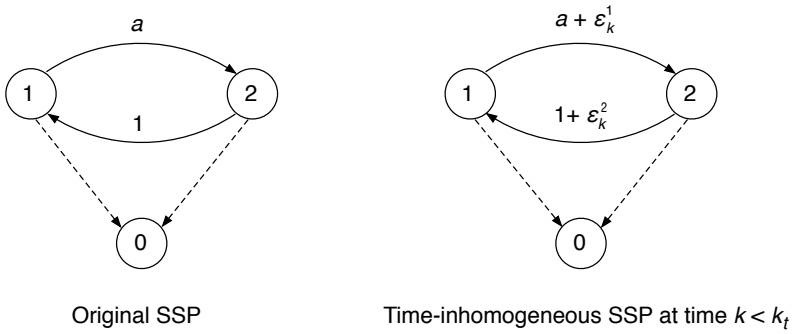

Figure 1: A 3-states example illustrating parts of the lower boundedness proof of Section 3.3. Transitions are deterministic. Control 0, indicated by the dashed lines, leads to the absorbing goal state 0. Control 1, indicated by the solid arcs, leads to a non-goal state with the expected transition cost indicated on the arc.

The state space is $S_o = \{0, 1, 2\}$. The feasible controls are $U(1) = U(2) = \{0, 1\}$, and all the transitions are deterministic. For control 1, $p_{12}(1) = p_{21}(1) = 1$ and the expected one-stage costs are

$$g(1, 1) = a \in (-1, 0], \qquad g(2, 1) = 1.$$

For control 0, $p_{10}(0) = p_{20}(0) = 1$ and the transition costs are zero. This SSP problem clearly satisfies Assumption 1.1. In the Q-learning algorithm, only two Q-factors, $Q_t(i, 1)$, $i = 1, 2$, are being updated, and the remaining Q-factors are fixed at zero. For simplicity we let $\gamma_t(i, 1) \in [0, 1]$ for all t.

The example is simple in that all the transitions are deterministic. Consequently, in the time-inhomogeneous SSP problems associated with the expressions of $\tilde{Q}_t(i, u)$ given by Lemma 3.3 (where time is indexed by k), the state transition probabilities $p_k^{iu,t}$ are time-invariant and identical to those in the original problem, and only the expected one-stage costs at states 1 and 2 vary over time – the variation is due to the randomness in the transition costs involved in the Q-learning algorithm. The right graph of Fig. 1 illustrates such a time-inhomogeneous SSP at time k: the expected one-stage costs of the SSP at states 1 and 2 have the form $a+\epsilon_k^1$ and $1+\epsilon_k^2$, respectively, and they vary within a δ-neighborhood of the original one-stage costs, for some $\delta > 0$ chosen in the construction of $\tilde{Q}_t$. Other quantities that can vary over time in the SSP problems associated with the expressions of $\tilde{Q}_t(1, 1)$ and $\tilde{Q}_t(2, 1)$, are the conditional probabilities of the randomized Markov policies $\{\nu_k^{iu,t}, k \geq 1\}$ for $i = 1, 2$ and $u = 1$.

For this example, any $\delta \leq (1 + a)/2$ is sufficiently small to fulfill the requirement of Lemma 3.8. Because with such δ, $a - \delta + 1 - \delta \geq 0$, and evidently no policy can have cost less than $a - \delta$ in an SSP whose expected one-stage costs vary within the intervals $[a - \delta, a + \delta]$ and $[1 - \delta, 1 + \delta]$ for states 1 and 2, respectively [cf. Fig. 1 (right)]. Consequently, $a - \delta$ is a lower bound of $\tilde{Q}_t(1, 1)$ and $\tilde{Q}_t(2, 1)$, $t \geq 0$, constructed in the proof for such δ.

We now do some direct calculation to illustrate the construction of $\{\tilde{Q}_t\}$ for this example. Consider a sample path on which

$$\tilde{g}_t(1, 1) \to a, \qquad \tilde{g}_t(2, 1) \to 1, \qquad \text{as } t \to \infty.$$

Let t_0 be such that

$$a - \delta \leq \tilde{g}_t(1, 1) \leq a + \delta, \qquad 1 - \delta \leq \tilde{g}_t(2, 1) \leq 1 + \delta, \qquad \forall\ t \geq t_0.$$

Let ν be the proper policy that applies control 0 at states 1 and 2:

$$\nu(0 \mid i) = 1, \qquad i = 1, 2.$$

Then the initial $\tilde{Q}_{t_0}$ is given by

$$\tilde{Q}_{t_0}(i,1) = \tilde{g}_{t_0}(i,1), \qquad i = 1,2,$$

(the other components of $\tilde{Q}_t$ are zero for all t), and they are the total costs of the policy ν for the initial state-control pairs $(i,1)$, in an SSP problem whose first-stage cost function is $\tilde{g}_{t_0}$ and whose one-stage cost functions for the remaining stages are g. For $t < t_0$, we have $\tilde{Q}_t = \tilde{Q}_{t_0}$ by definition.

For the purpose of illustration, let us assume that on the sample path, the Q-learning algorithm updates both Q-factors at time t_0 and updates only $Q(1,1)$ at time t_0+1, with these updates being

$$\begin{aligned}
Q_{t_0+1}(1,1) &= \big(1-\gamma_{t_0}(1,1)\big)Q_{t_0}(1,1) + \gamma_{t_0}(1,1)\big(a+\omega_{t_0}(1,1)+Q_{\tau_1}(2,1)\big),\\
Q_{t_0+1}(2,1) &= \big(1-\gamma_{t_0}(2,1)\big)Q_{t_0}(2,1) + \gamma_{t_0}(2,1)\big(1+\omega_{t_0}(2,1)+Q_{\tau_2}(1,1)\big),\\
Q_{t_0+2}(1,1) &= \big(1-\gamma_{t_0+1}(1,1)\big)Q_{t_0+1}(1,1) + \gamma_{t_0+1}(1,1)\big(a+\omega_{t_0+1}(1,1)+Q_{t_0+1}(2,1)\big),
\end{aligned}$$

where the stepsizes are in $(0,1]$ and $\tau_1, \tau_2 \le t_0$. We express the corresonding components of $\tilde{Q}_{t_0+1}$ and $\tilde{Q}_{t_0+2}$ in the form given in Lemma 3.3. By definition

$$\begin{aligned}
\tilde{Q}_{t_0+1}(1,1) &= \big(1-\gamma_{t_0}(1,1)\big)\,\tilde{Q}_{t_0}(1,1) + \gamma_{t_0}(1,1)\,\big(a+\omega_{t_0}(1,1)+\tilde{Q}_{\tau_1}(2,1)\big)\\
&= \big(1-\gamma_{t_0}(1,1)\big)\,\tilde{g}_{t_0}(1,1) + \gamma_{t_0}(1,1)\,\big(a+\omega_{t_0}(1,1)\big) + \gamma_{t_0}(1,1)\,\tilde{g}_{t_0}(2,1)\\
&= \tilde{g}_{t_0+1}(1,1) + \gamma_{t_0}(1,1)\,\tilde{g}_{t_0}(2,1), \qquad (4.1)
\end{aligned}$$

where the definition of $\tilde{g}_{t_0+1}(1,1)$ is used to obtain the last equality. Equation (4.1) shows that $\tilde{Q}_{t_0+1}(1,1)$ is equal to the cost of the Markov policy $\{\nu_k^{11,t_0+1}, k \ge 1\}$ for the initial state-control pair $(1,1)$, with

$$\nu_1^{11,t_0+1}(1 \mid 2) = \gamma_{t_0}(1,1), \qquad \nu_1^{11,t_0+1}(0 \mid 2) = 1-\gamma_{t_0}(1,1), \qquad (4.2)$$

$$\nu_1^{11,t_0+1}(\cdot \mid 1) = \nu_k^{11,t_0+1}(\cdot \mid i) = \nu(\cdot \mid i), \qquad i = 1,2,\ k \ge 2, \qquad (4.3)$$

in an SSP problem with time-varying one-stage cost functions $\{g_k^{11,t_0+1}, k \ge 0\}$, where the first- and second-stage cost functions are given by

$$g_0^{11,t_0+1}(i,1) = \tilde{g}_{t_0+1}(i,1), \qquad i = 1,2, \qquad (4.4)$$

$$g_1^{11,t_0+1}(1,1) = a, \qquad g_1^{11,t_0+1}(2,1) = \tilde{g}_{t_0}(2,1), \qquad (4.5)$$

and for the remaining stages, the one-stage cost functions are given by

$$g_k^{11,t_0+1}(1,1) = a, \qquad g_k^{11,t_0+1}(2,1) = 1, \qquad k \ge 2,$$

the same as the cost function of the original problem. (The transition probabilities are the same as in the original SSP problem and the one-stage costs for control 0 are all equal to 0.)

A similar calculation shows that

$$\tilde{Q}_{t_0+1}(2,1) = \tilde{g}_{t_0+1}(2,1) + \gamma_{t_0}(2,1)\,\tilde{g}_{t_0}(1,1), \qquad (4.6)$$

and it is equal to the cost of the Markov policy $\{\nu_k^{21,t_0+1}, k \ge 1\}$ for the initial state-control pair $(2,1)$, with

$$\begin{aligned}
&\nu_1^{21,t_0+1}(1 \mid 1) = \gamma_{t_0}(2,1), \qquad \nu_1^{21,t_0+1}(0 \mid 1) = 1-\gamma_{t_0}(2,1),\\
&\nu_1^{21,t_0+1}(\cdot \mid 2) = \nu_k^{21,t_0+1}(\cdot \mid i) = \nu(\cdot \mid i), \qquad i = 1,2,\ k \ge 2,
\end{aligned}$$

in an SSP problem with time-varying one-stage cost functions $\{g_k^{21,t_0+1}, k \ge 0\}$, where the first- and second-stage cost functions are given by

$$g_0^{21,t_0+1}(i,1) = \tilde{g}_{t_0+1}(i,1), \qquad i = 1,2,$$

$$g_1^{21,t_0+1}(1,1) = \tilde{g}_{t_0}(1,1), \qquad g_1^{21,t_0+1}(2,1) = 1,$$

and the remaining one-stage cost functions are the same as the cost function of the original problem.

Finally, for $\tilde{Q}_{t_0+2}(1,1)$, using its definition and the expressions of $\tilde{Q}_{t_0+1}(1,1)$ and $\tilde{Q}_{t_0+1}(2,1)$ in Eqs. (4.1), (4.6), and using also the definition of $\tilde{g}_{t_0+2}(1,1)$, we have

$$\begin{aligned}\tilde{Q}_{t_0+2}(1,1) &= \big(1-\gamma_{t_0+1}(1,1)\big)\tilde{Q}_{t_0+1}(1,1) + \gamma_{t_0+1}(1,1)\big(a+\omega_{t_0+1}(1,1)+\tilde{Q}_{t_0+1}(2,1)\big)\\ &= \tilde{g}_{t_0+2}(1,1) + \big(1-\gamma_{t_0+1}(1,1)\big)\,\gamma_{t_0}(1,1)\cdot\tilde{g}_{t_0}(2,1) + \gamma_{t_0+1}(1,1)\,\tilde{g}_{t_0+1}(2,1)\\ &\quad + \gamma_{t_0+1}(1,1)\,\gamma_{t_0}(2,1)\cdot\tilde{g}_{t_0}(1,1). \end{aligned} \tag{4.7}$$

Thus $\tilde{Q}_{t_0+2}(1,1)$ is equal to the cost of the Markov policy $\{\nu_k^{11,t_0+2}, k \geq 1\}$ for the initial state-control pair $(1,1)$ in an SSP problem with time-varying one-stage cost functions $\{g_k^{11,t_0+2}, k \geq 0\}$. Here the Markov policy is given by

$$\nu_1^{11,t_0+2}(1 \mid 2) = \big(1-\gamma_{t_0+1}(1,1)\big)\,\gamma_{t_0}(1,1) + \gamma_{t_0+1}(1,1), \tag{4.8}$$

$$\nu_1^{11,t_0+2}(0 \mid 2) = 1 - \nu_1^{11,t_0+2}(1 \mid 2),$$

$$\nu_2^{11,t_0+2}(1 \mid 1) = \gamma_{t_0+1}(1,1)\,\gamma_{t_0}(2,1), \tag{4.9}$$

$$\nu_2^{11,t_0+2}(0 \mid 1) = 1 - \nu_2^{11,t_0+2}(1 \mid 1),$$

with all the other unspecified components of ν_k^{11,t_0+2} being identical to those of the proper policy ν. In the SSP problem, for the first three stages, the one-stage cost functions are given by

$$g_0^{11,t_0+2}(i,1) = \tilde{g}_{t_0+2}(i,1), \quad i=1,2, \qquad g_1^{11,t_0+2}(1,1) = a, \tag{4.10}$$

$$g_1^{11,t_0+2}(2,1) = \frac{\big(1-\gamma_{t_0+1}(1,1)\big)\gamma_{t_0}(1,1)\cdot\tilde{g}_{t_0}(2,1) + \gamma_{t_0+1}(1,1)\tilde{g}_{t_0+1}(2,1)}{\big(1-\gamma_{t_0+1}(1,1)\big)\gamma_{t_0}(1,1) + \gamma_{t_0+1}(1,1)} \in [1-\delta, 1+\delta], \tag{4.11}$$

and

$$g_2^{11,t_0+2}(1,1) = \tilde{g}_{t_0}(1,1), \qquad g_2^{11,t_0+2}(2,1) = 1. \tag{4.12}$$

For the remaining stages, the one-stage cost functions are the same as the cost function of the original problem.

If we modify this example by introducing self-transitions at states 1 or 2, then the state transition probabilities of the above time-inhomogeneous SSP problems will also vary over time due to the simulation noise in Q-learning, and they can be calculated in the manner of the proof of Lemma 3.3. However, the preceding direct calculations are highly simplified due to the special nature of this example and illustrate only parts of the proof of Section 3.3. Even for SSP problems that are just slightly more complex, the full proof arguments of Section 3.3 become necessary, as the readers may verify on their example problems.

Acknowledgements

We thank Prof. John Tsitsiklis for mentioning to us the stepsize-truncation technique described in Section 2.3 and for suggesting that we consider the example of Section 4, as well as for other helpful discussion. This work was supported by the Air Force Grant FA9550-10-1-0412 and by NSF Grant ECCS-0801549.

References

[ABB02] J. Abounadi, D. P. Bertsekas, and V. Borkar, *Stochastic approximation for non-expansive maps: Application to Q-learning algorithms*, SIAM J. on Control and Optimization **41** (2002), 1–22.

[Bor08] V. S. Borkar, *Stochastic approximation: A dynamic viewpoint*, Hindustan Book Agency, New Delhi, 2008.

[BT91] D. P. Bertsekas and J. N. Tsitsiklis, *An analysis of stochastic shortest path problems*, Mathematics of Operations Research **16** (1991), no. 3, 580–595.

[BT96] ———, *Neuro-dynamic programming*, Athena Scientific, Belmont, MA, 1996.

[KY03] H. J. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications*, 2nd ed., Springer-Verlag, New York, 2003.

[Put94] M. L. Puterman, *Markov decision processes: Discrete stochastic dynamic programming*, John Wiley & Sons, New York, 1994.

[Sen81] E. Seneta, *Non-negative matrices and Markov chains*, 2nd ed., Springer-Verlag, New York, 1981.

[Tsi94] J. N. Tsitsiklis, *Asynchronous stochastic approximation and Q-learning*, Machine Learning **16** (1994), 185–202.

[Wat89] C. J. C. H. Watkins, *Learning from delayed rewards*, Ph.D. thesis, Cambridge University, England, 1989.

[YB11] H. Yu and D. P. Bertsekas, *Q-learning and policy iteration algorithms for stochastic shortest path problems*, Annals of Oper. Res. (2011), to appear.

[Yu11a] H. Yu, *Some proof details for asynchronous stochastic approximation algorithms*, 2011, on-line at: http://www.mit.edu/~janey_yu/note_asaproofs.pdf.

[Yu11b] ———, *Stochastic shortest path games and q-learning*, LIDS Report 2871, MIT, 2011.