Temporal-Difference Learning with State-Action-Dependent Discounting

Chien-Hsiang Yeh*

Australian National University

June 19, 2024

Abstract. This paper extends model-free learning algorithms, including Q-learning, SARSA, and double Q-learning to the learning with state-action-dependent discount factors. We allow the discount factor to be greater than one with positive probability, but the expected multiplicative of discount factors satisfies the "eventually discounting" condition in the sense that we replace $\beta < 1$ by $\rho(L) < 1$, where $\rho(L)$ denotes the spectral radius of an appropriate matrix dominating the expected discounted future value.

Keywords: State-action-dependent discounting, Eventual discounting, Q-learning, SARSA(0), double Q-learning

1. Introduction

Reinforcement learning algorithms, such as Q-learning introduced by Watkins (1989), learn the optimal actions from the interaction between agents and environments. In reinforcement learning, the discount factor is a meta-parameter of the performance of learning and is often set to be a less-than-one constant. However, in economics and finance, it is well-known that discount factors vary over time and are state-action-dependent discounting (Cochrane, 2011, Hills and Nakata, 2018). For example, asset pricing models consider stochastic discount factors, a function of risk preference, and

*Email: chien.yeh@anu.edu.au. The author thanks Prof. John Stachurski for the guidance and suggestions. The author also thanks Assoc. Prof. Timothy Kam and Prof. Fedor Iskhakov for the invaluable comments.

the current and future state-dependent consumption levels (Lucas Jr, 1978, Campbell and Ammer, 1993, Rosenberg and Engle, 2002, Cochrane, 2009, 2011, Hansen and Renault, 2010).

To apply Q-learning in economic or finance models, it is necessary to generalize the constant discounting to stochastic state-action-dependent discounting. In particular, we are interested in the case that the discount factor may be greater than one with a positive probability. It has been noted by Nakata (2016), Hills and Nakata (2018) and Hubmer et al. (2021) point out that if discount factors are defined as reciprocals of gross interest rate, they exhibit dynamic behavior and may occasionally exceed one.

In existing literature, Yoshida et al. (2013) show the convergence of Q-learning with state-dependent discount factors and provide a framework to optimize the state-dependent discount function, demonstrating superior performance compared to a constant discount factor in their simulation. Similarly, Sharma et al. (2021) explore the convergence of Q-learning and SARSA algorithms with both state and action-dependent discount factors. However, both of them assume that the discount factors are strictly less than one, whence their Bellman operator and Q-factor Bellman operator are evidently contraction maps.

To address this gap, this paper shows the convergence of three model-free reinforcement learning algorithms in finite Markov decision process under eventually discounting circumstances. These algorithms include the Q-learning, on-policy learning SARSA, and double Q-learning algorithms (Watkins and Dayan, 1992, Singh et al., 2000, Hasselt, 2010).

Following the assumptions outlined in Stachurski and Zhang (2021), we assume that the discount factors are eventually discounting: for all action policies σ there exists a $n_\sigma \in \mathbb{N}$ such that

$$\sup_x \mathbb{E}_x \prod_{t=0}^{n_\sigma - 1} \beta(X_t, \sigma(X_t), X_{t+1}) < 1,$$

where β is a function of states X_t and actions/policies $\sigma(X_{t+1})$ referred to as discount factors.[1] In words, the dynamics exhibit eventual discounting when the expected multiplicative of discount factors are eventually less than one for any policies.

[1] $\{X_t\}_{t \geqslant 0}$ is a Markov process defined in Section 2.1.

Our conditions do not rule out $\beta_t = \beta(X_t, \sigma(X_t), X_{t+1}) > 1$ with positive probability. On the other hand, the conventional proofs of the convergences for the Q-learning, SARSA, and double Q-learning rely on the fact that the discount factor is strictly less than one, whence the Bellman operator is contracting (Watkins and Dayan, 1992, Tsitsiklis, 1994, Bertsekas and Tsitsiklis, 1995, Singh et al., 2000, Hasselt, 2010).

We further analyze the contraction of the Bellman operator in the weighted norm of the eigenvector corresponding to the spectral radius. This further provides the convergence rates and error bounds for dynamic programming algorithms when there exists state-action-dependent. We use this fact to pin down the bound for optimal Q-factor.

Finally, we extend the Stochastic Approximation algorithm and Q-learning to cases with concave operators. Since a concave operator may have a unique fixed point and be globally stable, Stochastic Approximation with a concave operator should converge whenever it is bounded. This provides an alternative method to prove the convergence of Q-learning.

Related literature - Regarding action-dependent discounting, endogenous time preferences or Uzawa time preference are broadly studied in small open economies such as Obstfeld (1990), Mendoza (1991), Schmitt-Grohé and Uribe (2003), Vasilev (2022), Durdu et al. (2009).

The related literature in the Markov decision process with state- or action-dependent discount factors includes Wei and Guo (2011), Minjárez-Sosa (2015), Wu et al. (2015), Wu and Zhang (2016), Jasso-Fuentes et al. (2022). The literature about the theory of dynamic programming with state-dependent stochastic discounts in economics and finance includes Stachurski and Zhang (2021), Toda (2021), Sargent and Stachurski (2023), Toda (2023).

The applications of Q-learning in economics are as follows. Park and Ryu (2022) study suppliers' collusion behaviors concerning supply chain ethics and transparency with Q-learning agents. Through simulations, they show that suppliers tend to exhibit low levels of ethics and transparency. Neuneier (1997) utilize Q-learning to study the optimal asset allocation. Waltman and Kaymak (2008) use Q-learning to model the learning behavior of firms in repeated Cournot oligopoly games and find that Q-learning firms learn to collude with each other. Calvano et al. (2020) finds that Q-learning artificial intelligence autonomously learns to charge supracompetitive prices

in an oligopoly model of repeated price competition. Charpentier et al. (2021) conduct a comprehensive survey on the applications of reinforcement learning in economics and finance.

The paper is structured as follows. Section 2.1 introduces the framework and reinforcement learning algorithms. Section 3 presents the main results of the convergences. Section 5 discusses the stability of algorithms. Section 6 extends Stochastic Approximation and Q-learning to cases with concavity.

2. Q-learning, SARSA, and Double Q-learning

In this section, we introduce the Markov decision process and the model-free learning algorithms, including Q-learning, SARSA, and double Q-learning.

2.1. Background.

A finite *Markov decision process (MDP)* is a tuple $(\mathsf{X}, \mathsf{A}, \Gamma, \beta, P, r)$ satisfying that (i) X and A are finite state space and action space, respectively, (ii) Γ is a nonempty correspondence from $\mathsf{X} \to \mathsf{A}$, referred to as the feasible correspondence, which defines the feasible state-action pairs $\mathsf{G} \coloneqq \{(x,a) \in \mathsf{X} \times \mathsf{A} \colon a \in \Gamma(x)\}$, (iii) $\beta \colon \mathsf{G} \times \mathsf{X} \to \mathbb{R}_+$ is a discount factor function, (iv) $r \colon \mathsf{G} \times \mathsf{X} \to \mathbb{R}$ is the reward function, and (v) a stochastic kernel $P \colon \mathsf{G} \times \mathsf{X} \to \mathbb{R}_+$ satisfying $\sum_{x' \in \mathsf{X}} P(x,a,x') = 1$ for all $(x,a) \in \mathsf{G}$.

Given an MDP M, the set of feasible policies is

$$\Sigma \coloneqq \{\sigma \in \mathsf{A}^{\mathsf{X}} \colon \sigma(x) \in \Gamma(x), \quad \forall x \in \mathsf{X}\}.$$

An MDP is *recurrent or ergodic* if the Markov chain corresponding to every deterministic stationary policy $\sigma \in \Sigma$ consists of a single recurrent class.[2] If the MDP is recurrent, then for every policy $\sigma \in \Sigma$ the induced Markov chain will eventually visit every state; that is, every state is visited infinitely often.

To any stationary policy $\sigma \in \Sigma$ and initial state x, consider a trajectory process $X \coloneqq \{X_t\}_{t \in \mathbb{N}_0}$ taking values in state space X and controlled by policies $\sigma(X) \coloneqq \{\sigma(X_t)\}_{t \in \mathbb{N}_0}$ such that X_{t+1} is generated by $P(X_t, \sigma(X_t), \cdot)$ for all $t \in \mathbb{N}_0$. The processes X and $\sigma(X)$ are well-defined on some measurable space $(\Omega, \mathscr{F}, \mathbb{P})$ satisfying

[2] A state x is recurrent if it is eventually visited or returned to.

$\mathbb{P}(X_{t+1} = x'|\mathscr{F}_t) = P(X_t, \sigma(X_t), x')$ almost surely for any $x' \in \mathsf{X}$ and $t \in \mathbb{N}_0$, where $\mathscr{F}_t$ denotes the information set up to time t defined by

$$\mathscr{F}_t \coloneqq \sigma\{X_0, \ldots, X_t, a_0, \ldots, a_t, r_0, \ldots, r_{t-1}, \beta_0, \ldots, \beta_{t-1}\}.$$

Let r_t be the reward drawn from a fixed reward distribution $r : \mathsf{G} \times \mathsf{X} \to \mathbb{R}$ such that $\mathbb{E}[r_t|(x, a, x') = (X_t, \sigma(X_t), X_{t+1}) = r(x, a, x')$, where the conditional expectation of r_t is $r(x, a, x')$ conditioning on $X_t = x, \sigma(X_t) = a$ and being governed by underlying state transition $X_{t+1} = x'$. The σ-value function v_σ is defined by

$$v_\sigma(x) \coloneqq \mathbb{E}_x^\sigma \left[\sum_{t=0}^{\infty} \prod_{i=0}^{t-1} \beta_i r_t \right]$$

where $\beta_i \coloneqq \beta(X_i, \sigma(X_i), X_{i+1})$ for all $i \in \mathbb{N}_0$, $\mathbb{E}_x^\sigma$ denotes expectation conditioning on $\{x_0 = x\}$ with transition probability measure P, and $\prod_{i=0}^{-1} \beta_i \coloneqq 1$ by convention. The maximum total reward or *value function* is

$$v^*(x) \coloneqq \sup_\sigma v_\sigma(x) \qquad (x \in \mathsf{X}).$$

A policy $\sigma \in \Sigma$ is *optimal* if $v_\sigma = v^*$. The Bellman equation is

$$v(x) = \max_{a \in \Gamma(x)} \left\{ r(x, a) + \sum_{x' \in \mathsf{X}} \beta(x, a, x') v(x') P(x, a, x') \right\} \quad (x \in \mathsf{X})$$

where $r(x, a) \coloneqq \sum_{x' \in \mathsf{X}} r(x, a, x') P(x, a, x')$ for all $(x, a) \in \mathsf{G}$. Given $v \in \mathbb{R}^{\mathsf{X}}$, a policy $\sigma \in \Sigma$ is *v-greedy* if

$$\sigma(x) \in \arg\max_{a \in \Gamma(x)} \left\{ r(x, a) + \sum_{x' \in \mathsf{X}} \beta(x, a, x') v(x') P(x, a, x') \right\} \quad (x \in \mathsf{X}).$$

The *Bellman operator* is

$$Tv(x) \coloneqq \max_{a \in \Gamma(x)} \left\{ r(x, a) + \sum_{x' \in \mathsf{X}} \beta(x, a, x') v(x') P(x, a, x') \right\} \quad (x \in \mathsf{X}).$$

2.2. **Eventual Discounting.** We introduce the main assumptions imposed on state-action-dependent discount factors as follows. Define the expected multiplicative of discount factor d_n by

$$d_n \coloneqq \max_{\sigma \in \Sigma} \max_{x \in \mathsf{X}} \left\{ \mathbb{E}_x^\sigma \prod_{t=0}^{n-1} \beta(X_t, \sigma(X_t), X_{t+1}) \right\}. \tag{1}$$

We say that the MDP is *eventually discounting* if there is $n \in \mathbb{N}$ such that $d_n < 1$. In other words, the MDP is eventually discounting if the expected multiplicative of discount factors is eventually smaller than one for any policy.

We briefly discuss the sufficient conditions for eventual discounting. Denote function $B\colon \mathsf{G} \times \mathbb{R}^{\mathsf{X}} \to \mathbb{R}$ as

$$B(x,a,v) \coloneqq r(x,a) + \sum_{x' \in \mathsf{X}} P(x,a,x')\beta(x,a,x')v(x') \qquad ((x,a,v) \in \mathsf{G} \times \mathbb{R}^{\mathsf{X}}).$$

Suppose that there exists a $|\mathsf{X}| \times |\mathsf{X}|$ matrix L such that

$$|B(x,a,v) - B(x,a,w)| \leqslant \sum_{x' \in \mathsf{X}} L(x,x')|v(x') - w(x')| \qquad ((x,a) \in \mathsf{G}). \tag{2}$$

for all $v, w \in \mathbb{R}^{\mathsf{X}}$. If the spectral radius is $\rho(L) < 1$, we can show that it is eventually discounting. One possible assumption is to choose L to be

$$L_m(x,x') \coloneqq \max_{a \in \Gamma(x)} \beta(x,a,x')P(x,a,x') \qquad ((x,x') \in \mathsf{X}^2) \tag{3}$$

and assume $\rho(L_m) < 1$.

Lemma 2.1. *If L_m defined by (3) satisfies $\rho(L_m) < 1$, then there exists an $n \in \mathbb{N}$ such that $d_n < 1$.*

Given σ, define the $|\mathsf{X}| \times |\mathsf{X}|$ matrix L_σ by

$$L_\sigma(x,x') \coloneqq \beta(x,\sigma(x),x')P(x,\sigma(x),x')$$

for all $(x,x') \in \mathsf{X} \times \mathsf{X}$. Induction implies

$$\sum_{x'} L_\sigma^n(x,x') = \mathbb{E}_x^\sigma \prod_{t=0}^{n-1} \beta(x_t,\sigma(X_t),X_{t+1}) \qquad (x \in \mathsf{X}).$$

Then, we can show that $\rho(L_\sigma) < 1$ if and only if there is $n \in \mathbb{N}$ such that $d_n^\sigma < 1$, where

$$d_n^\sigma \coloneqq \max_x \mathbb{E}_x^\sigma \prod_{t=0}^{n-1} \beta(X_t,\sigma(X_t),X_{t+1}).$$

Moreover, (1) implies that there is $n \in \mathbb{N}$ such that

$$\max_\sigma d_n^\sigma = \max_\sigma \max_x \sum_{x'} L_\sigma^n(x,x') < 1.$$

Hence, we can further show that (1) is equivalent to

$$\max_\sigma \rho(L_\sigma) < 1.$$

Example 2.1. In economics and finance, discount factors are frequently determined as the reciprocals of gross real interest rates: $\beta_t = 1/(1+i_t)$, where $i_t \in \mathbb{R}$ is the real interest rate. It is well-known that real interest rates could be negative when there is zero lower bound for nominal interest rates (Hills and Nakata, 2018, Nakata, 2016, Hubmer et al., 2021). Hills et al. (2019) and Hubmer et al. (2021) consider an AR(1) process: $\beta_t = Z_t$, where $\{Z_t\}$ follows

$$Z_{t+1} = \rho_Z Z_t + (1-\rho_Z)\mu_Z + \sigma_\varepsilon \varepsilon_{t+1} \qquad \{\varepsilon_t\} \overset{IID}{\sim} N(0,1). \tag{4}$$

Hubmer et al. (2021) calibrate $\rho_Z = 0.992, \mu_Z = 0.944, \sigma_\varepsilon = 0.0006$. They discretize the process onto a grid of $N = 15$ states by Tauchen's method which allows us to write the operator L defined in (3) as

$$L_{ij} = \beta(x_i)P(x_i, x_j), \qquad 1 \leqslant i, j \leqslant N.$$

The spectral radius of matrix L is 0.9469 computed by Stachurski and Zhang (2021).

We can compute the value function and optimal policy by value function iteration, which is summarized below.

Proposition 2.1. *For an eventually discounting MDP,*

(i) the value function v^* is the unique solution to Bellman's equation in $\mathbb{R}^{\mathsf{X}}$,
(ii) $\lim_{k\to\infty} T^k v = v^*$ for all $v \in \mathbb{R}^{\mathsf{X}}$, and
(iii) a feasible policy is optimal if and only if it is v^*-greedy.

The proof can be found in Sargent and Stachurski (2023).

2.3. **Q-learning.** For each $v \in \mathbb{R}^{\mathsf{X}}$, the *Q-factor* corresponding to v is the function

$$Q(x,a) = r(x,a) + \sum_{x' \in \mathsf{X}} v(x')\beta(x,a,x')P(x,a,x') \quad (x,a) \in \mathsf{G}.$$

Denote Q^* as the Q-factor corresponding to v^*:

$$Q^*(x,a) = r(x,a) + \sum_{x' \in \mathsf{X}} v^*(x')\beta(x,a,x')P(x,a,x') \quad (x,a) \in \mathsf{G}.$$

The Bellman equation implies that

$$v^*(x) = \max_{a \in \Gamma(x)} Q^*(x,a) \qquad (x \in \mathsf{X}).$$

The goal of the Q-learning algorithm is to learn $Q^*(x,a)$ for all state-action pairs. Denote $\{x_t\}_{t\geqslant 0}$ as the realization of the Markov process where x_{t+1} is generated by $P(x_t,a_t,\cdot)$ for $t\in\mathbb{N}$, given $\{a_t\}_{t\geqslant 0}$. The Q-learning iterates the vector $Q_t\in\mathbb{R}^{\mathsf{G}}$ following the rule:

$$\begin{aligned} Q_{t+1}(x_t,a_t) =&(1-\alpha_t(x_t,a_t))Q_t(x_t,a_t) \\ &+\alpha_t(x_t,a_t)\left[r_t+\beta_t\max_{b\in\Gamma(x_{t+1})}Q_t(x_{t+1},b)\right], \end{aligned} \tag{5}$$

for $t\in\mathbb{N}_0$, where x_t, a_t, r_t, and α_t are the state, action, reward, and step size at time step t. In detail, given realized $\{(x_t,a_t)\}_{t\in\mathbb{N}-0}$, $\alpha_t(x,a)\in[0,1]$ is a step-size coefficient that $\alpha_t(x,a)=0$ for all $(x,a)\neq(x_t,a_t)$; that is, α_t is set to zero for those are outside of the support of $\{(x_t,a_t)\}_{t\in\mathbb{N}_0}$. Moreover, r_t is a random sample of reward that $\mathbb{E}[r_t|(x,a,x')=(x_t,a_t,x_{t+1})]=r(x,a,x')$, β_t is a random sample of discount factor that $\mathbb{E}[\beta_t|(x,a,x')=(x_t,a_t,x_{t+1})]=\beta(x,a,x')$, and x_{t+1} is a random successor state generated by $P(x_t,a_t,\cdot)$.

2.4. **SARSA.** We follow the terminology in Singh et al. (2000). The update rule for SARSA follows

$$\begin{aligned} Q_{t+1}(x_t,a_t) =&[1-\alpha_t(x_t,a_t)]Q_t(x_t,a_t) \\ &+\alpha_t(x_t,a_t)\left[r_t+\beta_t Q_t(x_{t+1},a_{t+1})\right], \end{aligned} \tag{6}$$

where a_{t+1} is determined by some learning policy introduced, and $\alpha_t(x,a)\in[0,1]$ with $\alpha_t(x,a)=0$ for $(x,a)\neq(x_t,a_t)$. Hence, SARSA is an on-policy algorithm, and its convergence depends on the learning policy.

A *learning policy* π is a set of probabilities $\Pr(\cdot|x,t,Q,n_t(x))$ such that action $a\in\mathsf{A}$ is selected with probability $\Pr(a|x,t,Q,n_t(x))$, given the history: state x, time step t, the current estimate Q of the optimal Q-value, the number of times, $n_t(x)$, that state x has been visited before time t. We say that a learning policy is *greedy* if it always selects the action that has the highest current Q-value.

A learning policy π is a *greedy-in-the-limit-with-infinite-exploration* (GLIE) learning policy if it is the set of probabilities $\Pr(a|x,t,Q,n_t(x))$ following two properties:

(a) each state is visited infinitely often, and each action is executed infinitely often in every state;

(b) in the limit, the learning policy is greedy with respect to the Q-value function with probability 1.

An example of a GLIE learning policy is ε-greedy exploration: at time step t in state x, picks the greedy action with probability $1-\varepsilon_t(x)$, and a random exploration action with probability $\varepsilon_t(x) = c/n_t(x)$ for $0 < c < 1$.

A learning policy is *restricted rank-based randomized (RRR)* if it selects actions probabilistically according to the ranks of Q-values: $\Pr(a|x,t,Q) = P^R(\rho(Q,x,a))$, where $\rho(Q,x,a)$ is the rank of action a in state x based on its action value $Q(x,a)$ for all a, and $P^R\colon \{1,\ldots,|A|\} \to \mathbb{R}$ maps action ranks to probabilities such that $P^R(1) \geqslant P^R(2) \geqslant \cdots \geqslant P^R(|A|)$ and $\sum_{i=1}^{|A|} P^R(i) = 1$. As an example, Q-learning is a SARSA iteration with an RRR learning policy that $P^R(1) = 1$ and $P^R(i) = 0$ for $i = 2,\ldots,|A|$.

We say that Q_t is generated by SARSA iteration ((6)) with a GLIE (RRR) learning policy π if a_{t+1} is selected by a GLIE (RRR) learning policy with $Q = Q_t$. The key difference between GLIE and RRR learning policies is that a GLIE learning policy has a decaying exploration such that the learning policy converges to the greedy policy over time, while an RRR learning policy has a persistent exploration. Therefore, Q_t generated by SARSA with a GLIE learning policy will converge to optimal Q-value Q^*, while Q_t may not converge to Q^* if the SARSA iteration follows an RRR learning policy.

Given a learning policy with ranking by Q-value, $\rho(Q,x,a)$, and ranking probability, P^R, let $\bar{Q}$ be the corresponding Q-factor satisfying

$$\bar{Q}(x,a) \coloneqq r(x,a) + \sum_{x' \in \mathsf{X}} P(x,a,x')\beta(x,a,x') \sum_{a' \in \mathsf{A}} P^R(\rho(\bar{Q},x',a'))\bar{Q}(x',a')$$

for $(x,a) \in G$. Given $\beta(x,a,x') \equiv \beta$ is constant, Singh et al. (2000) prove that Q_t value computed by SARSA with a GLIE learning policy converges to optimal Q^*, and Q_t updated with an RRR learning policy converges to the Q-factor function, $\bar{Q}$, corresponding to that learning policy.

On the other hand, the ranking policy can depend on (x,a) pair only without referencing Q-factor. Let $\Pi \coloneqq \{f\colon \mathsf{A} \to \{1,2,\ldots,|A|\}\colon f \text{ is a bijection}\}$ denote the set of permutations of A, where each $f \in \Pi$ ranks actions. A *restricted policy* $\bar{\pi}\colon \mathsf{X} \to \Pi$ ranks actions in each state without action value Q. That is, $\bar{\pi}(x,a) \coloneqq \bar{\pi}(x)(a)$ is the assigned rank of action a in state x. Given ranking probability P^R, $P^R(\bar{\pi}(x,a))$ is the probability that the policy selects action a in state x under restricted policy $\bar{\pi}$ that

ranks a. The Q-factor of a restricted policy $\bar{\pi}$, denoted $\bar{Q}^{\pi}$, satisfies

$$\bar{Q}^{\bar{\pi}}(x,a) = r(x,a) + \sum_{x' \in \mathsf{X}} P(x,a,x')\beta(x,a,x') \sum_{a' \in \mathsf{A}} P^R(\bar{\pi}(x',a'))\bar{Q}^{\bar{\pi}}(x',a')$$

for $(x,a) \in G$. We say a restricted policy $\bar{\pi}$ is *optimal* under probabilities of ranks P^R if

$$\bar{\pi} \in \arg\max_{\pi \in \Pi} \bar{Q}^{\pi}.$$

Also, the *greedy restricted policy* for a Q-value function Q is $\bar{\pi}(x,a) = \rho(Q,x,a)$, which ranks actions with their corresponding Q-values. Singh et al. (2000) shows that the greedy restricted policy with respect to $\bar{Q}$ (estimated by SARSA with an RRR learning policy) is an optimal restricted policy.

2.5. **Double Q-learning.** Given random variables $X_1, \ldots, X_n$, since $\mathbb{E}(\max_i\{X_i\}) \geqslant \max_i \mathbb{E}(X_i)$ by Jensen inequality, Q-learning is known to overestimate optimal values during experiments. That is, $\max_b Q_t(x_{t+1}, b)$ is an estimate for $\mathbb{E}\{\max_b Q_t(x_{t+1}, b)\}$, rather than for $\max_b \mathbb{E}[Q_t(x_{t+1}, b)]$ as desired, where the expectation is the average over all possible runs of the same experiments, and Q_t iteration is a sample mean that approximates Q^*. To avoid overestimation, Hasselt (2010) introduces double Q-learning, as outlined in Algorithm 1.

There are two Q-functions in double Q-learning: Q^A and Q^B. Each Q-function, say Q^A, is updated with another Q-value, Q^B, with Q^A-greedy action a^*. Since Q^A and Q^B update with different sets of experiment samples, Q^B is an unbiased estimate for Q-value at Q^A-greedy action. Hasselt (2010) illustrates that since $\mathbb{E}[Q^B(x', a^*)] \leqslant \max_a \mathbb{E}[Q^A(x', a)]$, double Q-learning may underestimate the action value.

Hasselt (2010) shows the convergence of double Q-learning assuming a constant discount factor $\beta < 1$. As illustrated in Algorithm 1, we assume that β is a random variable observed at the time that the Q-functions are updated. Moreover, β is governed by a parameterized function such that there exists eventual discounting.

3. Main Results

This section presents the assumptions and main results for the convergence of Q-learning, SARSA, and double Q-learning with state-action-dependent discount factors.

Algorithm 1: Double Q-learning

Initialize Q^A, Q^B, x
repeat
  Choose a, based on $Q^A(x,\cdot)$ and $Q^B(x,\cdot)$. Observe r, β, x'
  Choose (e.g. random) either UPDATE(A) or UPDATE(B)
  if UPDATE(A) **then**
    Define $a^* = \arg\max_a Q^A(x',a)$
    $Q^A(x,a) \leftarrow Q^A(x,a) + \alpha(x,a)\left(r + \beta Q^B(x',a^*) - Q^A(x,a)\right)$
  else if UPDATE(B) **then**
    Define $b^* = \arg\max_a Q^B(x',a)$
    $Q^B(x,a) \leftarrow Q^B(x,a) + \alpha(x,a)\left(r + \beta Q^A(x',b^*) - Q^B(x,a)\right)$
  end
  $x \leftarrow x'$
until end

3.1. **Assumptions and Convergences.** We first introduce the standard Robbins-Monro conditions for the stochastic approximation algorithm.

Let $\mathscr{F}_t$ be the information field up to time t:

$$\mathscr{F}_t = \sigma\{x_0, \ldots, x_t, a_0, \ldots, a_t, \alpha_0, \ldots, \alpha_t, Q_0, \ldots, Q_t, r_0, \ldots, r_{t-1}, \beta_0, \ldots, \beta_{t-1}\} \quad (7)$$

Let Ω be the sample space of all possible trajectories of $\{(x_t, a_t, r_t, \beta_t)\}_{t\in\mathbb{N}}$ and $\mathscr{F} = \bigotimes_{t\in\mathbb{N}_0} \mathscr{F}_t$. Let $\mathbb{P}$ be the probability measure on $(\Omega, \mathscr{F})$. For a trajectory $\omega \in \Omega$, define $\mathcal{T}_{x,a}(\omega) \subset \mathbb{N}$ be the set of times at which an update of $Q_t(x,a)$ is performed.

Assumption 3.1. The following conditions hold:

(a) X and A are finite;

(b) the stepsizes $\{\alpha_t\}_{t\in\mathbb{N}_0}$ is a sequence of random variables on $(\Omega, \mathbb{P}, \mathscr{F})$ such that $\alpha_t(x,a) \in [0,1]$, $\alpha_t(x,a) = 0$ for $t \notin \mathcal{T}_{x,a}(\omega)$ and

$$\sum_{t\in\mathcal{T}_{x,a}(\omega)} \alpha_t(x,a) = \infty, \qquad \sum_{t\in\mathcal{T}_{x,a}(\omega)} \alpha_t^2(x,a) < \infty$$

for all $(x,a) \in \mathsf{G}$ and $\mathbb{P}$-almost all $\omega \in \Omega$; and

Assumption 3.1 implies that each state-action pair will be visited infinitely many times by temporal-difference learning, and $\alpha_t(x,a) \to 0$ for each $(x,a) \in \mathsf{G}$. Next, we outline

the assumptions associated with state-action-dependent discounting. Define a *spectral radius* of a $|\mathsf{X}| \times |\mathsf{X}|$ matrix L by $\rho(L) \coloneqq \max\{|\lambda| \in \mathbb{R}^n \colon \lambda \text{ is an eigenvalue of } L\}$.

Assumption 3.2 (eventual-discounting)**.** For any $\sigma \in \Sigma$, $\rho(L_\sigma) < 1$, where $L_\sigma(x, x') \coloneqq \beta(x, \sigma(x), x')P(x, \sigma(x), x')$ for all $(x, x') \in \mathsf{X} \times \mathsf{X}$.

Assumption 3.3 (eventual-discounting)**.** There is a non-negative $|\mathsf{X}| \times |\mathsf{X}|$ matrix L such that $\beta(x, a, x')P(x, a, x') \leqslant L(x, x')$ for all $(x, a, x') \in \mathsf{G} \times \mathsf{X}$ and $\rho(L) < 1$.

Assumption 3.2 or 3.3 implies that the MDP is eventually discounting. Compared to a constant discount factor, Assumption 3.2 and 3.3 are more general in the sense that discount factors are states and actions dependent and can exceed one under some states and actions. Moreover, Assumption 3.3 is sufficient to Assumption 3.2. We can show that the policy operator or the Bellman operator is globally stable if either Assumption 3.2 or 3.3 holds. As discussed in Section 2.2, the assumptions imply that there is $n \in \mathbb{N}$ such that

$$\sup_{\sigma \in \Sigma} \sup_{x \in \mathsf{X}} \mathbb{E}_x^\sigma \prod_{t=0}^{n-1} \beta_t < 1$$

where $\beta_t = \beta(x_t, \sigma(x_t), x_{t+1})$ and $\mathbb{E}_x^\sigma$ denote the expectation conditioning on $x_0 = x$ and the transition probability follows $P(x, \sigma(x), x')$ for all $(x, x') \in \mathsf{X} \times \mathsf{X}$.

The convergences of Q-learning, SARSA, and double Q-learning are presented below with the above assumptions, in particular, with the state-action-dependent discount factors and eventual discounting.

Proposition 3.1. *If Assumption 3.1 holds, and either Assumption 3.2 or 3.3 holds, then $\{Q_t\}_{t \geqslant 0}$ generated by Q-learning algorithm (5) converges to Q^* w.p.*1*..*

Proposition 3.2. *If MDP is recurrent, Assumption 3.1 is satisfied, and either Assumption 3.2 or 3.3 holds, then the SARSA iterate $\{Q_t\}_{t \geqslant 0}$, generated by (6), with a GLIE learning policy π converges to Q^* w.p.*1*. and the corresponding learning policy π_t converges to the optimal policy σ^* w.p.*1*..*

Proposition 3.3. *If MDP is recurrent, Assumption 3.1 is satisfied, either Assumption 3.2 or 3.3 holds, and $\Pr(a_{t+1} = a | Q_t, x_{t+1}) = P^R(\rho(Q_t, x_{t+1}, a_{t+1}))$, then the SARSA iterate $\{Q_t\}_{t \geqslant 0}$, generated by (6), with an RRR learning policy converges to $\bar{Q}$ w.p.*1*. Moreover, the greedy restricted policy is the optimal restricted policy.*

Proposition 3.4. *If MDP is recurrent, Assumption 3.1 holds, and either Assumption 3.2 or 3.3 holds, and both Q^A and Q^B update infinitely often, then both $\{Q_t^A\}_{t\geqslant 0}$ and $\{Q_t^B\}_{t\geqslant 0}$ converge to optimal Q-value Q^* w.p.1.*

4. Proofs for Main Results

In this section, we provide the proofs for the propositions in Section 3.1.

4.1. **Preliminaries.** We denote $\|\cdot\|$ as the maximum norm $\|\cdot\|_\infty$. For any positive vector $w \in \mathbb{R}^n$, define the weighted maximum norm $\|\cdot\|_w$ by

$$\|v\|_w \coloneqq \max_x \frac{|v(x)|}{w(x)} \qquad (v \in \mathbb{R}^n).$$

A self-map F on U is *globally stable* if F has a unique fixed point u^* and $F^k u \to u^*$ for all $u \in U$.

4.2. **Proofs of Proposition 3.1.** We first show the convergence of Q-learning iteration. To apply stochastic approximation, define the Q-factor Bellman operator $H\colon \mathbb{R}^{\mathsf{G}} \to \mathbb{R}^{\mathsf{G}}$ by

$$\begin{aligned} HQ(x,a) &\coloneqq \mathbb{E}_{(x,a)} r(x,a,x') + \mathbb{E}_{(x,a)}\left[\beta(x,a,x') \max_{b\in\Gamma(x')} Q(x',b)\right] \\ &= B\left(x,a,\max_b Q(\cdot,b)\right) \qquad ((x,a)\in\mathsf{G}, Q\in\mathbb{R}^{\mathsf{G}}) \end{aligned}$$

We can verify that Q^* is the fixed point of H. The proof is completed by the following lemmas. We first show that H is a contraction map under some weighted norm. To simplify the notation, we define $\mathcal{M}q(x) \coloneqq \max_a q(x,a)$ for all $x \in \mathsf{X}$ and all $q \in \mathbb{R}^{\mathsf{G}}$ in the following lemmas.

Lemma 4.1. *If Assumption 3.2 holds, then there exists a positive vector $\varphi \in \mathbb{R}^{\mathsf{G}}$ and $\gamma < 1$ such that $\|HQ - HR\|_\varphi \leqslant \gamma\|Q-R\|_\varphi$ for all $Q, R \in \mathbb{R}^{\mathsf{G}}$.*

Proof. Let Assumption 3.2 hold. Proposition 2.1 implies that v^* is the unique fixed point to T. Define $Q^* \in \mathbb{R}^{\mathsf{G}}$ by

$$Q^*(x,a) = r(x,a) + \mathbb{E}_{x,a}\beta(x,a,X')v^*(X') \quad ((x,a)\in\mathsf{G}.)$$

Since $v^* = Tv^*$, we have

$$v^*(x) = \max_{a\in\Gamma(x')}\{r(x,a) + \mathbb{E}_{x,a}\beta(x,a,X')v^*(X')\} = \max_{a\in\Gamma(x)} Q^*(x,a),$$

for all $x \in \mathsf{X}$. It implies

$$\begin{aligned} HQ^*(x,a) &= r(x,a) + \mathbb{E}_{x,a}\left[\beta(x,a,X') \max_{a' \in \Gamma(X')} Q^*(X',a')\right] \\ &= r(x,a) + \mathbb{E}_{x,a}\beta(x,a,X')v^*(X') = Q^*(x,a) \end{aligned}$$

for all $(x,a) \in \mathsf{G}$. Hence, Q^* is the fixed point of H. Next, we construct the weighted vector φ by considering the constant reward $\hat{r}(x,a,x') \equiv -1$ for all $(x,a,x') \in \mathsf{G} \times \mathsf{X}$. Since there is a unique fixed point $\hat{Q} \in \mathbb{R}^{\mathsf{G}}$ when $\hat{r} = -\mathbb{1}$, we have[3]

$$\begin{aligned} \hat{Q}(x,a) = H\hat{Q}(x,a) &= \sum_{x'} P(x,a,x')[\hat{r}(x,a,x') + \beta(x,a,x') \max_{a'} \hat{Q}(x',a')] \\ &= -1 + \sum_{x'} P(x,a,x')\beta(x,a,x')\mathcal{M}\hat{Q}(x') \\ &\leqslant -1 + \max_{a \in \Gamma(x)} \sum_{x'} P(x,a,x')\beta(x,a,x')\mathcal{M}\hat{Q}(x') \qquad ((x,a) \in \mathsf{G}) \end{aligned}$$

Taking the maximum over A on the left, we obtain

$$\mathcal{M}\hat{Q}(x) \leqslant -1 + \max_{a \in \Gamma(x)} \sum_{x'} P(x,a,x')\beta(x,a,x')\mathcal{M}\hat{Q}(x') \qquad (x \in \mathsf{X}.) \tag{8}$$

Observe that $\mathcal{M}\hat{Q}$ is the optimal value, and there exists an optimal policy $\hat{\sigma} \in \Sigma$ such that $\mathcal{M}\hat{Q} = v_{\hat{\sigma}}$. Since $\rho(L_{\hat{\sigma}}) < 1$, we have $L_{\hat{\sigma}}^n \mathbb{1} \to 0$ as $n \to \infty$, and $T_{\hat{\sigma}}$ is globally stable. Then, the iteration of $v_{\hat{\sigma}} = T_{\hat{\sigma}}^n v_{\hat{\sigma}} = (I + L_{\hat{\sigma}} + \cdots + L_{\hat{\sigma}}^n)(-\mathbb{1})$ for all $n \in \mathbb{N}$ converges and yields

$$\mathcal{M}\hat{Q}(x) = v_{\hat{\sigma}}(x) = \lim_{n\to\infty} \mathbb{E}_x^{\hat{\sigma}} \sum_{t=0}^{n} \prod_{i=0}^{t-1} \beta(X_i, \hat{\sigma}(X_i), X_{i+1})(-1) \leqslant -1$$

for all $x \in \mathsf{X}$. Hence, we have $\hat{v} \leqslant -\mathbb{1}$. Let $\varphi = -\mathcal{M}\hat{Q} \geqslant \mathbb{1}$. Hence, (8) implies

$$\max_{a \in \Gamma(x)} \sum_{x'} P(x,a,x')\beta(x,a,x')\varphi(x') \leqslant \varphi(x) - 1 \leqslant \varphi(x) \max_{y} \frac{\varphi(y) - 1}{\varphi(y)} = \gamma\varphi(x),$$

[3]Denote $\mathbb{1} \in \mathbb{R}^{\mathsf{X}}$ as the vector of ones.

where $\gamma \coloneqq \max_x\{(\varphi(x)-1)/\varphi(x)\} < 1$. Now, for all $Q, R \in \mathbb{R}^{\mathsf{G}}$, we obtain

$$\begin{aligned}
|HQ(x,a) - HR(x,a)| &= \left|\sum_{x'} \beta(x,a,x')P(x,a,x')(\max_{a'} Q(x',a') - \max_{a'} R(x',a'))\right| \\
&\leqslant \sum_{x'} \beta(x,a,x')P(x,a,x') \left|\max_{a'} Q(x',a') - \max_{a'} R(x',a')\right| \\
&\leqslant \sum_{x'} \beta(x,a,x')P(x,a,x') \max_{a'} |Q(x',a') - R(x',a')| \\
&\leqslant \max_{a\in\Gamma(x)} \sum_{x'} \beta(x,a,x')P(x,a,x') \max_{a'} |Q(x',a') - R(x',a')| \\
&\leqslant \max_{a\in\Gamma(x)} \sum_{x'} \beta(x,a,x')P(x,a,x')\varphi(x') \max_{y} \frac{\mathcal{M}(|Q-R|)(y)}{\varphi(y)} \\
&= \|Q-R\|_\varphi \max_{a\in\Gamma(x)} \sum_{x'} \beta(x,a,x')P(x,a,x')\varphi(x') \\
&\leqslant \|Q-R\|_\varphi \gamma\varphi(x)
\end{aligned} \tag{9}$$

for all $(x,a) \in \mathsf{G}$. Dividing $\varphi(x)$ and taking supremum over G, we obtain the contraction $\|HQ - HR\|_\varphi \leqslant \gamma\|Q-R\|_\varphi$ for all $Q, R \in \mathbb{R}^{\mathsf{G}}$. □

To this end, Lemma 4.1 implies that H is globally stable and has a unique fixed point if Assumption 3.2 holds. Since 3.3 implies Assumption 3.2, H is also contractive on some weighted supremum norm. Alternatively, we can show that H is eventually contracting by the property of L, which bounds the expected discount factors.

Lemma 4.2. *If Assumption 3.3 holds, then operator H has a unique fixed point and is globally stable.*

Proof. Consider an eventually discounted MDP. We first show that there exists $k \in \mathbb{N}$ such that H^k is a contraction map. Let $Q, R \in \mathbb{R}^{\mathsf{G}}$. Since the MDP is eventually discounting, we obtain

$$\begin{aligned}
|HQ(x,a) - HR(x,a)| &= |B(x,a,\max_{a'} Q(\cdot,a')) - B(x,a,\max_{a'} R(\cdot,a'))| \\
&\leqslant \sum_{x'\in\mathsf{X}} L(x,x') \left|\max_{a'} Q(x',a') - \max_{a'} R(x',a')\right| \\
&\leqslant \sum_{x'\in\mathsf{X}} L(x,x') \max_{a'} |Q(x',a') - R(x',a')|
\end{aligned}$$

for all $(x,a) \in \mathsf{G}$. Taking the maximum over A on the left, we obtain

$$\max_a |HQ(x,a) - HR(x,a)| \leqslant \sum_{x' \in \mathsf{X}} L(x,x') \max_{a'} |Q(x',a') - R(x',a')|$$

for all $(x,a) \in \mathsf{G}$. Hence, we have $\mathcal{M}(|HQ - HR|) \leqslant L(\mathcal{M}|Q - R|)$. Observe that $\mathcal{M}|H^2Q - H^2R| \leqslant L(\mathcal{M}|HQ - HR|) \leqslant LL(\mathcal{M}|Q - R|) = L^2(\mathcal{M}|Q - R|)$. Hence, the induction shows that

$$\begin{aligned}\mathcal{M}(|H^jQ - H^jR|) &\leqslant L^j(\mathcal{M}|Q - R|) \\ &\leqslant \|L^j\|\|Q - R\|.\end{aligned}$$

Taking the maximum over X on the left, we obtain $\|H^jQ - H^jR\| \leqslant \|L^j\|\|Q - R\|$. Since $\rho(L) < 1$, the Gelfand's formula implies that there exists $k \in \mathbb{N}$ such that $\|L^k\| < 1$. To this end, H^k is a contraction map. The Banach Contraction Mapping Theorem implies that H^k has a unique fixed point Q^* in $\mathbb{R}^{\mathsf{G}}$ and $Q^* = \lim_{j\to\infty} H^{jk}Q$ for any $Q \in \mathbb{R}^{\mathsf{G}}$. To see that Q^* is the unique fixed point of H, since H^k is globally stable, we can fix $\varepsilon > 0$ and choose $j > 0$ such that

$$\|H^{jk}(HQ^*) - Q^*\| < \varepsilon.$$

This implies that $\|HH^{jk}Q^* - Q^*\| = \|HQ^* - Q^*\| < \varepsilon$. Since this holds for all $\varepsilon > 0$, we obtain $\|HQ^* - Q^*\| = 0$ so that Q^* is also a fixed point of H. Finally, by the argument

$$\lim_{m\to\infty} H^mQ = \lim_{\substack{s\to\infty \\ m=sk+t; s,t\in\mathbb{N}}} H^{sk+t}Q = \lim_{\substack{s\to\infty \\ m=sk+t; s,t\in\mathbb{N}}} H^{sk}(H^tQ) = Q^*,$$

H is globally stable, and Q^* is the unique fixed point to H. □

Lemma 4.3. *If Assumption 3.3 holds, then there exists a positive vector $\varphi \in \mathbb{R}^{\mathsf{G}}$ and $\gamma < 1$ such that $\|HQ - HR\|_\varphi \leqslant \gamma\|Q - R\|_\varphi$ for all $Q, R \in \mathbb{R}^{\mathsf{G}}$.*

Proof. Suppose Assumption 3.3 holds. Then, Lemma 4.2 implies that H has a unique fixed point. We construct the weighted vector φ by considering the reward $\hat{r}(x,a,x') \equiv -1$ for all $(x,a,x') \in \mathsf{G} \times \mathsf{X}$. Since there is a unique fixed point $\hat{Q} \in \mathbb{R}^{\mathsf{G}}$

when $\hat{r} = -\mathbb{1}$, we have

$$\begin{aligned}\hat{Q}(x,a) = H\hat{Q}(x,a) &= \sum_{x'} P(x,a,x')[\hat{r}(x,a,x') + \beta(x,a,x') \max_{a'} \hat{Q}(x',a')] \\ &= -1 + \sum_{x'} P(x,a,x')\beta(x,a,x') \max_{a'} \hat{Q}(x',a') \\ &\leqslant -1 + \sum_{x'} L(x,x')\mathcal{M}\hat{Q}(x') \qquad ((x,a) \in \mathsf{G})\end{aligned} \tag{10}$$

Taking the maximum over A on the left, we obtain

$$\mathcal{M}\hat{Q}(x) \leqslant -\mathbb{1} + \sum_{x'} L(x,x')\mathcal{M}\hat{Q}(x') \qquad (x \in \mathsf{X}.)$$

Let $\hat{\varphi} = -\mathcal{M}\hat{Q}$. Then, since $\rho(L) < 1$ and L is non-negative, we have

$$\mathcal{M}\hat{Q} \leqslant (I - L)^{-1}(-\mathbb{1}) = (I + L + L^2 + \dots)(-\mathbb{1}) \leqslant -\mathbb{1}.$$

Hence, (10) implies

$$\sum_{x'} L(x,x')\varphi(x') \leqslant \varphi(x) - 1 \leqslant \varphi(x) \max_{y} \frac{\varphi(y) - 1}{\varphi(y)} = \gamma\varphi(x),$$

where $\gamma \coloneqq \max_x\{(\varphi(x) - 1)/\varphi(x)\} < 1$. Now, for all $Q, R \in \mathbb{R}^{\mathsf{G}}$, we obtain

$$\begin{aligned}|HQ(x,a) - HR(x,a)| &= \left|\sum_{x'} \beta(x,a,x')P(x,a,x'))(\max_{a'} Q(x',a') - \max_{a'} R(x',a'))\right| \\ &\leqslant \sum_{x'} L(x,x')\mathcal{M}|Q - R|(x') \\ &\leqslant \left(\sum_{x'} L(x,x')\varphi(x')\right)\left(\max_{y} \frac{\mathcal{M}|Q - R|(y)}{\varphi(y)}\right) \\ &\leqslant \gamma\varphi(x) \max_{y} \frac{\mathcal{M}|Q - R|(y)}{\varphi(y)} \\ &= \gamma\varphi(x)\|Q - R\|_{\varphi} \qquad ((x,a) \in \mathsf{G})\end{aligned}$$

for all $(x,a) \in \mathsf{G}$. Dividing $\varphi(x,a)$ on both sides and taking the maximum on the left, we obtain the contraction $\|HQ - HR\|_{\varphi} \leqslant \gamma\|Q - R\|_{\varphi}$ for all $Q, R \in \mathbb{R}^{\mathsf{G}}$. □

The intuition of contraction in the weighted supremum norm for Q-learning is discussed as follows. Rewriting Q-learning iteration (5), we have

$$Q_{t+1}(x_t,a_t) = [1 - \alpha_t(x_t,a_t)]Q_t(x_t,a_t) + \alpha_t(x_t,a_t)\left[HQ_t(x_t,a_t) + w_t(x_t,a_t)\right],$$

where

$$w_t(x,a) = r_t - \mathbb{E}_{(x,a)} r_t + \max_{b \in \Gamma(x_{t+1})} Q_t(x_{t+1}, b) - \mathbb{E}_{(x,a)} \left[\max_{b \in \Gamma(X')} Q_t(X', b) \right].$$

It is known that if H is a contraction map, then Q_t converges to the fixed point of H, i.e., Q^* (Tsitsiklis, 1994). Now, observe that the Q-learning iteration (5) is equivalent to

$$\begin{aligned} \frac{Q_{t+1}(x_t,a_t)}{\varphi(x_t,a_t)} = [1 - \alpha_t(x_t,a_t)] \frac{Q_t(x_t,a_t)}{\varphi(x_t,a_t)} \\ + \alpha_t(x_t,a_t) \left[\frac{H[\varphi(x_t,a_t)\frac{Q_t(x_t,a_t)}{\varphi(x_t,a_t)}]}{\varphi(x_t,a_t)} + \frac{w_t(x_t,a_t)}{\varphi(x_t,a_t)} \right], \end{aligned} \tag{11}$$

Therefore, it is equivalent to iterate $q_t(x,a) \coloneqq Q_t(x,a)/\varphi(x,a)$ with respect to the map $\tilde{H}$ defined by

$$\tilde{H}q \coloneqq \Phi^{-1} H(\Phi q) \qquad (q \in \mathbb{R}^{\mathsf{G}}).$$

where $\Phi q(x,a) \coloneqq \varphi(x,a) q(x,a)$ and $\Phi^{-1} q(x,a) = q(x,a)/\varphi(x,a)$ for all $(x,a) \in \mathsf{G}$ and $q \in \mathbb{R}^{\mathsf{G}}$. In the above lemmas, we have $\varphi(x,a) \equiv \varphi(x)$. If H is a contracting map in $\|\cdot\|_\varphi$ norm, then $\tilde{H}$ is a contraction in maximum norm.

We use the following lemma of stochastic approximation to prove convergence of algorithms. The methodology follows Jaakkola et al. (1993), Tsitsiklis (1994), Singh et al. (2000), Melo (2001), and Hasselt (2010).

Lemma 4.4 (Singh et al. (2000))**.** *Consider a stochastic process* $\{(\alpha_t, \Delta_t, F_t)\}_{t \geqslant 0}$ *such that* $\alpha_t, \Delta_t, F_t : X \to \mathbb{R}$ *and*

$$\Delta_{t+1}(x) = (1 - \alpha_t(x))\Delta_t(x) + \alpha_t(x) F_t(x)$$

Let $\mathscr{F}_t$ *be a sequence of increasing* σ*-fields such that* α_0, Δ_0 *are* $\mathscr{F}_0$*-measurable and* α_t, Δ_t *and* F_{t-1} *are* $\mathscr{F}_t$*-measurable for* $t \geqslant 1$*. Assume that the following statements hold:*

(a) the set of possible states X *is finite;*
(b) $0 \leqslant \alpha_t(x) \leqslant 1$ *,* $\sum_t \alpha_t(x) = \infty$*,* $\sum_t \alpha_t^2(x) < \infty$ *w.p.*1*;*
(c) $\|\mathbb{E}[F_t(\cdot)|\mathscr{F}_t]\|_w \leqslant \kappa \|\Delta_t\|_w + c_t$*, where* $\kappa \in (0,1)$ *and* c_t *converges to zero w.p.*1*;*
(d) $\mathrm{Var}[F_t(x)|\mathscr{F}_t] \leqslant K(1 + \|\Delta_t\|_w)^2$ *, where* K *is some constant.*

Then, $\Delta_t \to 0$ *w.p.*1*.*

proof of Proposition 3.1. Suppose that all assumptions in the statement hold. Let $\Delta_t(x,a) = Q_t(x,a) - Q^*(x,a)$ for all $(x,a) \in \mathsf{G}$ and rearrange (5) as

$$\begin{aligned}\Delta_{t+1}(x_t,a_t) =& (1-\alpha_t(x_t,a_t))\Delta_t(x_t,a_t) \\ &+ \alpha_t(x_t,a_t)\left[r_t + \beta_t \max_{b\in\Gamma(x_{t+1})} Q_t(x_{t+1},b) - Q^*(x_t,a_t)\right] \\ =& (1-\alpha_t(x_t,a_t))\Delta_t(x_t,a_t) + \alpha_t(x_t,a_t)F_t(x_t,a_t),\end{aligned}$$

where we write

$$F_t(x_t,a_t) = r_t + \beta_t \max_{b\in\Gamma(x_{t+1})} Q_t(x_{t+1},b) - Q^*(x_t,a_t).$$

We let $F_t(x,a) = 0$ if $(x,a) \neq (x_t,a_t)$. Here, x_{t+1} is a random sample generated from $P(x_t,a_t,\cdot)$ for all $t \in \mathbb{N}_0$. Let $\mathscr{F}_t = \sigma\{Q_0, x_0, a_0, \alpha_0, r_0, \beta_0, \ldots, x_t, a_t, \alpha_t, r_{t-1}, \beta_{t-1}\}$ be the history up to time t. Then, Δ_t, α_t and F_{t-1} are $\mathscr{F}_t$-measurable.

To apply Lemma 4.4, we need to show (1) $\|\mathbb{E}[F_t(\cdot)|\mathscr{F}_t\|_w \leqslant \kappa\|\Delta_t\|_w + c_t$, for some $\kappa \in (0,1)$ and c_t converges to zero w.p.1, and (2) $\mathrm{Var}[F_t(x)|\mathscr{F}_t] \leqslant K(1+\|\Delta_t\|_w)^2$ for some constant K. From the setup of F_t, we have

$$\begin{aligned}\mathbb{E}[F_t(x,a)|\mathscr{F}_t] &= \sum_{x'\in\mathsf{X}} P(x,a,x')\left[r(x,a,x') + \beta(x,a,x')\max_{b\in\Gamma(x')} Q_t(x',b) - Q^*(x,a)\right] \\ &= HQ_t(x,a) - Q^*(x,a)\end{aligned}$$

Since $HQ^* = Q^*$, Lemma 4.1 or Lemma 4.3 implies

$$\|\mathbb{E}[F_t(\cdot,\cdot)|\mathscr{F}_t]\|_\varphi = \|HQ_t - HQ^*\|_\varphi \leqslant \gamma\|Q_t - Q^*\|_\varphi = \gamma\|\Delta_t\|_\varphi. \tag{12}$$

for some $0 < \gamma < 1$ and positive vector φ. Moreover, since X and A are finite, we have $\mathrm{Var}(r_t|\mathscr{F}_t) = \mathrm{Var}(r(x_t,a_t,X')) < \infty$ and $\mathrm{Var}(\beta_t|\mathscr{F}_t) = \mathrm{Var}(\beta(x_t,a_t,X')) < \infty$. It implies that there exists $C \in \mathbb{R}$ such that

$$\begin{aligned}\mathrm{Var}[F_t(x,a)|\mathscr{F}_t] &= \mathbb{E}\left[\left(r_t + \beta_t \max_{b\in\Gamma(X')} Q_t(X',b) - Q^*(x,a) - (HQ_t(x,a) - Q^*(x,a))\right)^2\right] \\ &= \mathbb{E}\left[\left(r_t + \beta_t \max_{b\in\Gamma(X')} Q_t(X',b) - HQ_t(x,a)\right)^2\right] \\ &= \mathrm{Var}\left[r_t + \beta_t \max_{b\in\Gamma(X')} Q_t(X',b)\right] \\ &\leqslant C(1+\|\Delta_t\|_\varphi)^2.\end{aligned}$$

Then, Lemma 4.4 shows that Δ_t converges to zero w.p.1, so Q_t converges to Q^* w.p.1. □

4.3. **Remainging Proofs in Section 3.1.** For the stability of SARSA, we can show that Q_t computed by SARSA iteration is bounded w.p.1 by Theorem 1 of Tsitsiklis (1994). Alternatively, note that the Q-value from Q-learning is an upper bound for Q values of SARSA. Moreover, consider a Q-learning process with min instead of max in the update rule:

$$\begin{aligned} Q_{t+1}(x_t,a_t) =& [1-\alpha_t(x_t,a_t)]Q_t(x_t,a_t) \\ &+ \alpha_t(x_t,a_t)\left[r_t + \beta_t \min_{b\in\Gamma(x_{t+1})} Q_t(x_{t+1},b)\right]. \end{aligned} \tag{13}$$

Clearly, Q-values from (13) iteration are the lower bounds for the Q-values of SARSA. Since update rule (13) is equivalent to the negative Q_t of the Q-learning (5) replacing r_t with $-r_t$, it also converges w.p.1.[4]

proof of Proposition 3.2. Suppose that Assumption 3.1 and 3.3 hold. Let $\Delta_t(x,a) = Q_t(x,a) - Q^*(x,a)$ for all $(x,a) \in \mathsf{G}$. The SARSA iterate becomes

$$\Delta_{t+1}(x_t,a_t) = (1-\alpha_t(x_t,a_t))\Delta_t(x_t,a_t) + \alpha_t(x_t,a_t)F_t(x_t,a_t),$$

where

$$\begin{aligned} F_t(x_t,a_t) =& r_t - Q^*(x_t,a_t) + \beta_t Q_t(x_{t+1},a_{t+1}) \\ =& r_t + \beta_t \max_{b\in\Gamma(x_{t+1})} Q_t(x_{t+1},b) - Q^*(x_t,a_t) + \beta_t\left[Q_t(x_{t+1},a_{t+1}) - \max_{b\in\Gamma(x_{t+1})} Q_t(x_{t+1},b)\right] \\ :=& F_t^Q(x_t,a_t) + \beta_t\left[Q_t(x_{t+1},a_{t+1}) - \max_{b\in\Gamma(x_{t+1})} Q_t(x_{t+1},b)\right] \\ :=& F_t^Q(x_t,a_t) + C_t(x_{t+1},a_{t+1}). \end{aligned}$$

Define $F_t(x,a) = F_t^Q(x,a) = C_t(x,a) = 0$ if $(x,a) \neq (x_t,a_t)$. Let

$$\mathscr{F}_t = \sigma\{Q_0, x_0, a_0, \alpha_0, r_0, \beta_0, \ldots, x_t, a_t, \alpha_t, r_{t-1}, \beta_{t-1}\}$$

be the history up to time t. Then, Δ_t, α_t and F_{t-1} are $\mathscr{F}_t$-measurable. Since $\mathbb{E}[F_t^Q(x,a)|\mathscr{F}_t] = HQ_t(x,a) - Q^*(x,a)$, it follows from (12) that

$$\|\mathbb{E}[F_t^Q(\cdot,\cdot)|\mathscr{F}_t]\|_\varphi \leqslant \gamma\|\Delta_t\|_\varphi,$$

[4]We can also prove the convergence of the iteration (13) by the same arguments in the proofs of Q-learning.

where $\gamma < 1$ and φ are obtained from Lemma 4.1 or Lemma 4.3. Therefore,

$$\begin{aligned}\|\mathbb{E}[F_t(\cdot,\cdot)|\mathscr{F}_t]\|_\varphi &\leqslant \|\mathbb{E}[F_t^Q(\cdot,\cdot)|\mathscr{F}_t]\|_\varphi + \|\mathbb{E}[C_t(\cdot,\cdot)|\mathscr{F}_t]\|_\varphi \\ &\leqslant \gamma\|\Delta_t\|_\varphi + \|\mathbb{E}[C_t(\cdot,\cdot)|\mathscr{F}_t]\|_\varphi\end{aligned}$$

Now, $\|\mathbb{E}[C_t(\cdot,\cdot)|\mathscr{F}_t]\|_\varphi$ converges to zero since $Q_t(x,a)$ stays bounded as discussed at the beginning of this subsection, and a GLIE learning policy converges to the optimal policy, whence C_t converges to zero w.p.1. Finally, since $\mathrm{Var}(r_t|\mathscr{F}_t) < \infty$ and $\mathrm{Var}(\beta_t|\mathscr{F}_t) < \infty$ by the finiteness of X and A, we have $\mathrm{Var}(F_t|\mathscr{F}_t) \leqslant C(1+\|\Delta_t\|_\varphi)^2$ for some constant C. Therefore, Lemma 4.4 concludes that Δ_t converges to zero w.p.1, and then Q_t converges to Q^* w.p.1. □

proof of Proposition 3.3. Let the conditions of the statements hold. Let $\Delta_t(x,a) \coloneqq Q_t(x,a) - \bar{Q}(x,a)$ for $(x,a) \in \mathsf{G}$, where $\bar{Q}$ is the Q-factor corresponding to the RRR learning policy. Rewriting (6), we have

$$\Delta_{t+1}(x_t,a_t) = (1-\alpha_t(x_t,a_t))\Delta_t(x_t,a_t) + \alpha_t(x_t,a_t)F_t(x_t,a_t)$$

where

$$F_t(x_t,a_t) \coloneqq r_t + \beta_t Q_t(x_{t+1},a_{t+1}) - \bar{Q}(x_t,a_t)$$

for $(x,a) \in \mathsf{G}$. Define $F_t(x,a) = 0$ if $(x,a) \neq (x_t,a_t)$ and let

$$\mathscr{F}_t = \sigma\{Q_0, x_0, a_0, \alpha_0, r_0, \beta_0, \ldots, x_t, a_t, \alpha_t, r_{t-1}, \beta_{t-1}\}.$$

Recall that $\bar{Q}$ function is

$$\bar{Q}(x,a) = r(x,a) + \sum_{x'\in\mathsf{X}} P(x,a,x')\beta(x,a,x') \sum_{a'\in\mathsf{A}} P^R(\rho(\bar{Q},x',a'))\bar{Q}(x',a') \quad (x,a) \in \mathsf{G}$$

Define operator $S\colon \mathbb{R}^{\mathsf{G}} \to \mathbb{R}^{\mathsf{G}}$ by $SQ(x,a) = \sum_{a\in\mathsf{A}} P^R(\rho(Q,x,a))Q(x,a)$ for $(x,a) \in \mathsf{G}$ and $Q \in \mathbb{R}^{\mathsf{G}}$. Note that for any $Q, Q' \in \mathbb{R}^{\mathsf{G}}$ we have

$$|SQ(x,a) - SQ'(x,a)| \leqslant \max_a |Q(x,a) - Q'(x,a)| \qquad ((x,a) \in \mathsf{G}).^{5}$$

[5] See Singh et al. (2000) Appendix C.

Suppose that Assumption 3.3 holds. Let $\gamma < 1$ and φ be defined as Lemma 4.3. Then, since a_{t+1} follows the RRR learning policy, we have the expectation

$$\begin{aligned}
|\mathbb{E}[F_t(x_t,a_t)|\mathscr{F}_t]| &= |\mathbb{E}[r_t + \beta_t Q_t(x_{t+1},a_{t+1}) - \bar{Q}(x_t,a_t)|\mathscr{F}_t]| \\
&= \left| \sum_{x' \in \mathsf{X}} P(x_t,a_t,x')\beta(x_t,a_t,x') S Q_t(x',a') - \sum_{x' \in \mathsf{X}} P(x_t,a_t,x')\beta(x_t,a_t,x') S\bar{Q}(x',a') \right| \\
&\leqslant \sum_{x' \in \mathsf{X}} P(x_t,a_t,x')\beta(x_t,a_t,x') \left| S Q_t(x',a') - S\bar{Q}(x',a') \right| \\
&\leqslant \sum_{x' \in \mathsf{X}} P(x_t,a_t,x')\beta(x_t,a_t,x') \max_{a'} |Q_t(x',a') - \bar{Q}(x',a')| \\
&\leqslant \sum_{x' \in \mathsf{X}} L(x_t,x') \max_{a'} |Q_t(x',a') - \bar{Q}(x',a')| \\
&\leqslant \sum_{x' \in \mathsf{X}} L(x_t,x')\varphi(x') \|Q_t - \bar{Q}\|_\varphi \leqslant \gamma\varphi(x_t)\|Q_t - \bar{Q}\|_\varphi
\end{aligned}$$

where the second equality follows from $\mathbb{E}(r_t|\mathscr{F}_t) = r(x_t,a_t)$ and $\Pr(a_{t+1} = a|Q_t, x_{t+1}) = P^R(\rho(Q_t,x_t,a))$, the third inequality follows from the assumption of eventual discounting, and the last two inequalities follow Assumption 3.3 and Lemma 4.3. On the other hand, suppose that Assumption 3.2 holds. Let γ and φ be defined as Lemma 4.1. Similar to the inequality (9) in Lemma 4.1, we have

$$\begin{aligned}
|\mathbb{E}[F_t(x_t,a_t)|\mathscr{F}_t]| &= |\mathbb{E}[r_t + \beta_t Q_t(x_{t+1},a_{t+1}) - \bar{Q}(x_t,a_t)|\mathscr{F}_t]| \\
&\leqslant \sum_{x' \in \mathsf{X}} P(x_t,a_t,x')\beta(x_t,a_t,x') \left| S Q_t(x',a') - S\bar{Q}(x',a') \right| \\
&\leqslant \max_a \sum_{x' \in \mathsf{X}} P(x_t,a_t,x')\beta(x_t,a_t,x') \max_{a'} |Q_t(x',a') - \bar{Q}(x',a')| \\
&\leqslant \max_a \sum_{x' \in \mathsf{X}} P(x_t,a_t,x')\beta(x_t,a_t,x')\varphi(x') \|Q_t - \bar{Q}\|_\varphi \\
&\leqslant \gamma\varphi(x_t)\|Q_t - \bar{Q}\|_\varphi
\end{aligned}$$

Dividing $\varphi(x_t)$ and taking supremum to either one of the above inequalities, we have $\|\mathbb{E}[F_t(\cdot,\cdot)|\mathscr{F}_t]\|_\varphi \leqslant \gamma\|Q_t - \bar{Q}\|_\varphi$. Now, Lemma 4.4 shows that Δ_t converges to zero w.p.1, so Q_t converges to $\bar{Q}$ w.p.1. The statement that the greedy restricted policy is optimal under P^R ranking strategy follows from Theorem 3 of Singh et al. (2000), where we consider the discounting function $\beta(x,a,x')$ instead of a constant β. □

proof of Proposition 3.4. Suppose that all conditions of the statement are satisfied. By symmetry, it suffices to show that Q^A converges to Q^*. To apply Lemma 4.4, let $\Delta_t \coloneqq Q_t^A - Q^*$. Denote $\alpha_t = \alpha_t(x_t, a_t)$, $a^* \coloneqq \arg\max_a Q^A(x_{t+1}, a)$ and $b^* \coloneqq \arg\max_b Q^B(x_{t+1}, b)$. We then have

$$\begin{aligned}\Delta_t(x_t, a_t) &= (1-\alpha_t)\Delta_t(x_t, a_t) + \alpha_t[r_t + \beta_t Q_t^B(x_{t+1}, a^*) - Q^*(x_t, a_t)] \\ &\coloneqq (1-\alpha_t)\Delta_t(x_t, a_t) + \alpha_t F_t(x_t, a_t)\end{aligned} \tag{14}$$

where

$$\begin{aligned}F_t(x_t, a_t) &\coloneqq r_t + \beta_t Q_t^B(x_{t+1}, a^*) - Q^*(x_t, a_t) \\ &= r_t + \beta_t Q_t^A(x_{t+1}, a^*) - Q^*(x_t, a_t) + \beta_t[Q_t^B(x_{t+1}, a^*) - Q_t^A(x_{t+1}, a^*)] \\ &\coloneqq F_t^Q(x_t, a_t) + \beta_t[Q_t^B(x_{t+1}, a^*) - Q_t^A(x_{t+1}, a^*)] \\ &\coloneqq F_t^Q(x_t, a_t) + C_t(x_t, a_t).\end{aligned}$$

Let $F_t(x, a) = F_t^Q(x, a) = C_t(x, a) = 0$ if $(x, a) \neq (x_t, a_t)$. Recall that $\mathbb{E}[F_t^Q(x_t, a_t)|\mathscr{F}_t] = HQ_t^A(x_t, a_t) - HQ^*(x_t, a_t)$, so we have

$$\|\mathbb{E}[F_t^Q(\cdot, \cdot)|\mathscr{F}_t]\|_\varphi \leqslant \gamma \|\Delta_t\|_\varphi,$$

where $\gamma < 1$ and φ are defined in Lemma 4.1 or Lemma 4.3. Since $\mathrm{Var}(r_t|\mathscr{F}_t) < \infty$ and $\mathrm{Var}(\beta_t|\mathscr{F}_t) < \infty$, the variance condition of Lemma 4.4 is satisfied. Therefore, it suffices to show that $C_t(x_t, a_t) = \beta_t(Q_t^B(x_{t+1}, a^*) - Q_t^A(x_{t+1}, a^*))$ converges to zero w.p.1.

Next, we show that $\Delta_t^{BA} \coloneqq Q_t^B - Q_t^A$ converges to zero w.p.1 by Lemma 4.4. Double Q-learning algorithm 1 implies that the update Δ_t follows

if Q^B is updated:

$$\begin{aligned}\Delta_{t+1}^{BA}(x_t, a_t) =& Q_{t+1}^B(x_t, a_t) - Q_{t+1}^A(x_t, a_t) \\ =& (1-\alpha_t)Q_t^B(x_t, a_t) + \alpha_t(r_t + \beta_t Q^A(x_{t+1}, b^*)) - Q_t^A(x_t, a_t) \\ =& (1-\alpha_t)\Delta_t^{BA}(x_t, a_t) + \alpha_t[r_t + \beta_t Q_t^A(x_{t+1}, b^*) - Q_t^A(x_t, a_t)]\end{aligned}$$

if Q^A is updated:

$$\begin{aligned}\Delta_{t+1}^{BA}(x_t, a_t) =& Q_{t+1}^B(x_t, a_t) - Q_{t+1}^A(x_t, a_t) \\ =& Q_t^B - [(1-\alpha_t)Q_t^A(x_t, a_t) + \alpha_t(r_t + \beta_t Q^B(x_{t+1}, a^*))] \\ =& (1-\alpha_t)\Delta_t^{BA}(x_t, a_t) - \alpha_t[r_t + \beta_t Q_t^B(x_{t+1}, a^*) - Q_t^B(x_t, a_t)].\end{aligned}$$

Suppose that the probabilities of updating Q^A and Q^B are equal, and the selection of updating Q^A or Q^B is independent of the sample. Then, we have

$$\Delta_{t+1}^{BA}(x_t, a_t) = (1 - \alpha_t(x_t, a_t))\Delta_t^{BA}(x_t, a_t) + \alpha_t(x_t, a_t)\tilde{F}_t(x, a), \tag{15}$$

where

$$\tilde{F}_t(x, a) = \begin{cases} r_t + \beta_t Q_t^A(x_{t+1}, b^*) - Q_t^A(x_t, a_t), & \text{w.p.} 1/2 \\ -r_t - \beta_t Q_t^B(x_{t+1}, a^*) + Q_t^B(x_t, a_t), & \text{w.p.} 1/2. \end{cases}$$

To use Lemma 4.4 for $\{\Delta_t^{BA}\}$ process, we need to show $\|\mathbb{E}[\tilde{F}_t(\cdot, \cdot)|\mathscr{F}_t]\| \leqslant \lambda \|\Delta_t^{BA}\|$ for some $\lambda < 1$. From the definition of $\tilde{F}_t$, we obtain

$$\begin{aligned} \mathbb{E}[\tilde{F}_t(x, a)|\mathscr{F}_t] &= \frac{1}{2}\left[\mathbb{E}(r_t + \beta_t Q_t^A(x_{t+1}, b^*)|\mathscr{F}_t) - Q_t^A(x_t, a_t)\right] \\ &\quad + \frac{1}{2}\left[-\mathbb{E}(r_t + \beta_t Q_t^B(x_{t+1}, a^*)|\mathscr{F}_t) + Q_t^B(x_t, a_t)\right] \\ &= \frac{1}{2}\Delta_t^{BA}(x_t, a_t) + \frac{1}{2}\mathbb{E}[\beta_t(Q_t^A(x_{t+1}, b^*) - Q_t^B(x_{t+1}, a^*))|\mathscr{F}_t]. \end{aligned}$$

Assume $\mathbb{E}[\beta_t Q_t^A(x_{t+1}, b^*)|\mathscr{F}_t] \geqslant \mathbb{E}[\beta_t Q_t^B(x_{t+1}, a^*)|\mathscr{F}_t]$. Suppose that Assumption 3.3 holds. Let $\gamma < 1$ and φ be defined as Lemma 4.3. Since the definition of a^* implies $Q_t^A(x_{t+1}, a^*) \geqslant Q_t^A(x_{t+1}, b^*)$, we have

$$\begin{aligned} |\mathbb{E}\{\beta_t[Q_t^A(x_{t+1}, b^*) - Q_t^B(x_{t+1}, a^*)]|\mathscr{F}_t\}| &= \mathbb{E}\{\beta_t[Q_t^A(x_{t+1}, b^*) - Q_t^B(x_{t+1}, a^*)]|\mathscr{F}_t\} \\ &\leqslant \mathbb{E}\{\beta_t[Q_t^A(x_{t+1}, a^*) - Q_t^B(x_{t+1}, a^*)]|\mathscr{F}_t\} \\ &= \sum_{x'} P(x_t, a_t, x')\beta(x_t, a_t, x')[Q_t^A(x', a^*) - Q_t^B(x', a^*)] \\ &\leqslant \sum_{x'} L(x_t, x') \max_{a'} |Q_t^A(x', a') - Q_t^B(x', a')| \\ &\leqslant \sum_{x'} L(x_t, x')\varphi(x')\|Q_t^A - Q_t^B\|_\varphi \\ &\leqslant \gamma\varphi(x_t)\|Q_t^A - Q_t^B\|_\varphi \end{aligned}$$

Dividing both sides by $\varphi(x_t)$ and taking supremum, we have

$$\|\mathbb{E}[\beta_t(Q_t^B(x_{t+1}, \cdot) - Q_t^A(x_{t+1}, \cdot))|\mathscr{F}_t]\|_\varphi \leqslant \gamma\|\Delta_t^{BA}\|_\varphi. \tag{16}$$

On the other hand, suppose that Assumption 3.2 holds. Let γ and φ be defined as Lemma 4.1. The inequality (9) in Lemma 4.1 implies

$$\begin{aligned}
|\mathbb{E}\{\beta_t[Q_t^A(x_{t+1},b^*)-Q_t^B(x_{t+1},a^*)]|\mathscr{F}_t\}| &= \mathbb{E}\{\beta_t[Q_t^A(x_{t+1},b^*)-Q_t^B(x_{t+1},a^*)]|\mathscr{F}_t\} \\
&\leqslant \mathbb{E}\{\beta_t[Q_t^A(x_{t+1},a^*)-Q_t^B(x_{t+1},a^*)]|\mathscr{F}_t\} \\
&= \sum_{x'} P(x_t,a_t,x')\beta(x_t,a_t,x')[Q_t^A(x',a^*)-Q_t^B(x',a^*)] \\
&\leqslant \max_a \sum_{x'} P(x_t,a_t,x')\beta(x_t,a_t,x') \max_{a'} |Q_t^A(x',a')-Q_t^B(x',a')| \\
&\leqslant \max_a \sum_{x'} P(x_t,a_t,x')\beta(x_t,a_t,x')\varphi(x')\|Q_t^A-Q_t^B\|_\varphi \\
&\leqslant \gamma\varphi(x_t)\|Q_t^A-Q_t^B\|_\varphi.
\end{aligned}$$

To this end, (16) holds if Assumption 3.2 holds.

Alternatively, if we assume $\mathbb{E}[\beta_t Q_t^A(x_{t+1},b^*)|\mathscr{F}_t] < \mathbb{E}[\beta_t Q_t^B(x_{t+1},a^*)|\mathscr{F}_t]$, then by the fact that $Q_t^B(x_{t+1},b^*) \geqslant Q_t^B(x_{t+1},a^*)$, the above argument also implies (16). Therefore, we have

$$\begin{aligned}
|\mathbb{E}[\tilde{F}_t(x,a)|\mathscr{F}_t]| &= \left|\frac{1}{2}\Delta_t^{BA}(x_t,a_t)+\frac{1}{2}\mathbb{E}[\beta_t(Q_t^A(x_{t+1},b^*)-Q_t^B(x_{t+1},a^*))|\mathscr{F}_t]\right| \\
&\leqslant \frac{1+\gamma}{2}\|\Delta_t^{BA}\|_\varphi
\end{aligned}$$

where $(1+\gamma)/2 < 1$. Since $\mathrm{Var}(r_t|\mathscr{F}_t) < \infty$ and $\mathrm{Var}(\beta_t|\mathscr{F}_t) < \infty$, we obtain $\mathrm{Var}(\tilde{F}_t|\mathscr{F}_t) \leqslant K(1+\|\Delta_t^{BA}\|)^2$ for some constant K. Then, Lemma 4.4 and (15) yield $\Delta_t^{BA} \to 0$ w.p.1. Therefore, C_t also converges to zero w.p.1. Finally, Lemma 4.4 shows that the origin process (14) converges: $\Delta_t = Q^A - Q^* \to 0$ w.p.1. □

5. Stability of Q-learning

The Bellman operator of an eventually discounting MDP can be viewed as a contraction map under some weighted norm. We find that the spectral radius $\rho(L)$ and its corresponding eigenvector determine the convergence of the value function iteration and other algorithms of such MDP. With this fact, we can determine the bound for optimal Q-value Q^*.

Define a policy operator $T_\sigma\colon \mathbb{R}^{\mathsf{X}} \to \mathbb{R}^{\mathsf{X}}$ as

$$T_\sigma v(x) \coloneqq r(x,\sigma(x)) + \sum_{x'} P(x,\sigma(x),a')\beta(x,\sigma(x),x')v(x') \qquad (v \in \mathbb{R}^{\mathsf{X}}, x \in \mathsf{X}).$$

The Bellman operator can be written as $Tv(x) = \max_{\sigma\in\Sigma} T_\sigma v(x)$ for all $x \in \mathsf{X}$. Assume further that L is irreducible such that L has an eigenvector φ corresponding to $\rho(L)$: $L\varphi = \rho(L)\varphi$. We define the maps:

$$\tilde{T}_\sigma v \coloneqq \Phi^{-1} T_\sigma(\Phi v); \quad \tilde{T} v \coloneqq \Phi^{-1} T(\Phi v); \qquad (v \in \mathbb{R}^{\mathsf{X}})$$
$$\tilde{H} q \coloneqq \hat{\Phi}^{-1} H(\hat{\Phi} q), \qquad (q \in \mathbb{R}^{\mathsf{G}})$$

where $\Phi = \operatorname{diag}(\varphi)$, $\hat{\Phi} = \operatorname{diag}(\hat{\varphi})$, and $\hat{\varphi}(x,a) \coloneqq \varphi(x)$ for all $(x,a) \in \mathsf{G}$. Observe that $\tilde{v}_\sigma$, $\tilde{v}$ and $\tilde{q}$ are the fixed points of $\tilde{T}_\sigma, \tilde{T}$ and $\tilde{H}$, respectively, if and only if $v_\sigma = \Phi\tilde{v}_\sigma$, $v^* = \Phi\tilde{v}$ and $Q^* = \hat{\Phi}\tilde{q}$ are the fixed points of T_σ, T and H, respectively. We show that $\tilde{T}_\sigma, \tilde{T}$, and $\tilde{H}$ are contraction maps with modulus $\rho(L)$, which further give the convergence rates of $\tilde{T}_\sigma, \tilde{T}$ and $\tilde{H}$ in the maximum norm, or the convergence rates of T_σ, T and H in $\|\cdot\|_\varphi$.

Lemma 5.1. *If Assumption 3.3 holds, and L is irreducible, then T_σ, T, and H are contraction maps in $\|\cdot\|_\varphi$ with modulus $\rho(L)$, where φ is the eigenvector of L corresponding to eigenvalue $\rho(L)$.*

Corollary 5.1. *If Assumption 3.3 holds and L is irreducible, then $\tilde{T}_\sigma, \tilde{T}$ and $\tilde{H}$ are contraction maps with modulus $\rho(L)$ under maximum norm.*

Therefore, we see that both T and H are globally stable, and we can analyze the convergence rates. For example, we have the following convergence rate and error bound for value function iteration (see, e.g., Bertsekas (2022) for the proof).]

Lemma 5.2. *If Assumption 3.3 holds, and L is irreducible, then*

(a) $\|T^k v - v^*\|_\varphi \leqslant \rho(L)^k \|v - v^*\|_\varphi$
(b) $\|T^{k+1} v - v^*\|_\varphi \leqslant \gamma \|T^{k+1} v - T^k v\|_\varphi$, *where* $\gamma = \rho(L)/(1-\rho(L))$.

Next, we use the contraction of H operator to find the boundedness of optimal Q^*. It also indicates the stability of the Q-learning in the sense that the expectation $\mathbb{E}(Q_{t+1}(x,a)|\mathscr{F}_t)$ is bounded if Q_t is bounded.

Proposition 5.1. *Suppose that Assumption 3.3 holds, L is irreducible, and $|r_t| \leqslant \bar{r}$ for all t w.p.1. Let Q_t be the Q-learning iteration. Then, the following statements are true.*

(a) $\|Q^*\|_\varphi \leqslant \|\bar{r}\|_\varphi/(1-\rho(L))$,

(b) if $\|Q_t\|_\varphi \leqslant \|\bar{r}\|_\varphi/(1-\rho(L))$*, then* $\|\mathbb{E}(Q_{t+1}(\cdot,\cdot)|\mathscr{F}_t)\|_\varphi \leqslant \|\bar{r}\|_\varphi/(1-\rho(L))$.

Proposition 5.1 shows that the Q-learning iteration is stable in the sense that Q_t is always expected to be bounded by $\|\bar{r}\|_\varphi/(1-\rho(L))$ for all t. Finally, note that all the results also hold if Assumption 3.2 holds, by Lemma 4.1.

6. Learning with Concavity

In this part, we explore cases where the Q-factor Bellman operator exhibits concavity. Initially, we show that if the operator in Stochastic Approximation, which has a desired fixed point, is concave in a bounded interval, then the Stochastic Approximation iteration converges to the fixed point of that operator, provided that iterations remain within the same interval. We next apply the result to Q-learning by assuming that the Q-factor Bellman operator is concave.

6.1. **Stochastic Approximation with Concavity.** Let $H\colon \mathbb{R}^n \to \mathbb{R}^n$ be a map on $\mathbb{R}^n$ such that $Hx = (Hx(1), \ldots, Hx(n))$ for all $x \in \mathbb{R}^n$. We are interested in computing x such that $Hx = x$. *Stochastic approximation* algorithm consists of updates of a vector $x \in \mathbb{R}^n$ with noisy for solving the fixed point of H. Let $\mathcal{T}_i \subset \mathbb{N}$ be the set of times at which an update of $x(i)$ is performed for $i \in \{1, \ldots, n\}$. The Stochastic Approximation iterates

$$x_{t+1} = \begin{cases} x_t(i), & \text{if } t \notin \mathcal{T}_i \\ (1-\alpha_t(i))x_t(i) + \alpha_t(i)(Hx_t(i) + w_t(i)), & \text{if } t \in \mathcal{T}_i \end{cases} \tag{17}$$

for all $i \in \{1, \ldots, n\}$, where $\alpha_t(i) \in [0,1]$ is a step size parameter, $w_t(i)$ is a random noise, and $x_0 \in \mathbb{R}^n$.

Let $\mathscr{F}_t$ be the σ-field of the algorithm information up to and including the point at which the step-size $\alpha_t(i)$ is selected, but just before the noise or update direction is determined. Specifically, we let

$$\mathscr{F}_t = \sigma\{x_0, \ldots, x_t, w_0, \ldots, w_{t-1}, \alpha_0, \ldots, \alpha_t\}.$$

Let Ω be the sample space of all possible trajectories of $\{(x_t, \alpha_t, w_t)\}$ and $\mathscr{F} = \bigotimes_{t\in\mathbb{N}_0} \mathscr{F}_t$. Let $\mathbb{P}$ be the probability measure on $(\Omega, \mathscr{F})$. The following two assumptions to stepsizes and noises are standard for Stochastic Approximation.

Assumption 6.1. The stepsizes $\{\alpha_t\}_{t\in\mathbb{N}_0}$ are a sequence of random variables defined on $(\Omega, \mathscr{F}, \mathbb{P})$ such that $\alpha_t(i) \in [0,1]$ and $\alpha_t(i) = 0$ for $t \in \mathcal{T}_i$ for all i and t. Moreover, we have

$$\sum_{t\in\mathcal{T}_i(\omega)} \alpha_t(i) = \infty, \text{ and } \sum_{t\in\mathcal{T}_i(\omega)} \alpha_t^2(i) < \infty$$

for all i and $\mathbb{P}$-almost all $\omega \in \Omega$.

Assumption 6.2.

(a) $\mathbb{E}[w_t(i)|\mathscr{F}_t] = 0$ for all i and t.

(b) There exist constants A and B such that for all i and t

$$\mathbb{E}[w_t^2(i)|\mathscr{F}_t] \leqslant A + B\|x_t\|^2$$

We assume that H is concave on some interval where it is globally stable and has a unique fixed point.

Assumption 6.3.

(a) H is increasing and concave on $[u, v] \subset \mathbb{R}^n$ with $u < v$,

(b) $Hv \leqslant v$, and

(c) there exists an $\varepsilon > 0$ such that $Hu \geqslant u + \varepsilon(v - u)$.

Theorem 6.1. *If Assumption 6.1, 6.2 and 6.3 hold, and $\{x_t\}$ generated by (17) is in $[u, v]$ with probability 1, then x_t converges to x^* with probability 1.*

6.2. **Q-learning with Concavity.** Let $c\colon \mathsf{G} \times \mathsf{X} \to \mathbb{R}_+$ denote a cost function. Suppose that the (future) value is adjusted by some concave function $\varphi\colon \mathbb{R} \to \mathbb{R}$ before taking expectation such that the Bellman equation becomes

$$v(x) = \min_{a\in\Gamma(x)} \mathbb{E}_{x,a}\varphi\left(c(x, a, X') + \beta(x, a, X')v(X')\right). \tag{18}$$

Let v^* be a solution to (18) and define $Q^*(x, a) \coloneqq \varphi(c(x, a, X') + \beta(x, a, X')v^*(X'))$ for all $(x, a) \in \mathsf{G}$. The corresponding Q-learning iteration follows

$$\begin{aligned} Q_{t+1}(x_t, a_t) = (1-&\alpha_t(x_t, a_t))Q_t(x_t, a_t) \\ &+ \alpha_t(x_t, a_t)\left[\varphi\left(c_t + \beta_t \min_{b\in\Gamma(x_{t+1})} Q_t(x_{t+1}, b)\right)\right] \end{aligned} \tag{19}$$

where $\alpha_t(x,a) \in [0,1]$ for all $(x,a) \in \mathsf{G}$, c_t and β_t satisfy $\mathbb{E}[c_t|(x,a',x') = (x_t,a_t,x_{t+1})] = c(x,a,x')$ and $\mathbb{E}[\beta_t|(x,a,x') = (x_t,a_t,x_{t+1})] = \beta(x,a,x')$, and x_{t+1} is a random successor state generated by $P(x_t,a_t,\cdot)$, given $Q_0 \in \mathbb{R}^{\mathsf{G}}$. Define the Q-factor Bellman operator $H\colon \mathbb{R}^{\mathsf{G}} \to \mathbb{R}^{\mathsf{G}}$ by

$$HQ(x,a) \coloneqq \mathbb{E}_{x,a}\varphi\left(c(x,a,X') + \beta(x,a,X') \min_{b\in\Gamma(X')} Q(X',b)\right) \tag{20}$$

for all $(x,a) \in \mathsf{G}$ and $Q \in \mathbb{R}^{\mathsf{G}}$.

Assumption 6.4.

(a) Both c and β are positive everywhere and bounded.

(b) $\varphi\colon \mathbb{R} \to \mathbb{R}$ is increasing, concave, and $\varphi(c(x,a,x')) > 0$ for all $(x,a,x') \in \mathsf{G}\times\mathsf{X}$.

(c) There is $K > 0$ such that $\varphi(\|c\| + \|\beta\|K) \leqslant K$.

Assumption 6.4 guarantees that $H(K\mathbb{1}) \leqslant K\mathbb{1}$. It also implies that $H0$ is everywhere positive and then there exists an $\varepsilon > 0$ such that $H0 \geqslant \varepsilon K\mathbb{1}$. To this end, H satisfies Assumption 6.3.

Example 6.1. This example demonstrates a standard Q-learning. If $\beta \in (0,1)$, φ is an identity map, and c is positive everywhere and bounded. Suppose that H is defined by

$$HQ(x,a) = c(x,a) + \beta\mathbb{E}_{x,a} \min_{b\in\Gamma(x')} Q(x',b) \qquad ((x,a) \in \mathsf{G}, Q \in \mathbb{R}^{\mathsf{G}}).$$

Then, we have

$$\begin{aligned} H\left(\frac{\|c\|}{1-\beta}\mathbb{1}\right)(x,a) &= c(x,a) + \beta\mathbb{E}_{x,a} \min_{a'}\left\{\frac{\|c\|}{1-\beta}\mathbb{1}(x',a')\right\} \\ &\leqslant \|c\| + \beta\frac{\|c\|}{1-\beta} = \frac{\|c\|}{1-\beta}. \end{aligned}$$

Therefore, H is a self-map on $[u,v]$, where $u(x,a) \equiv 0$ and $v(x,a) \equiv \|c\|/(1-\beta))$ for $(x,a) \in \mathsf{G}$. Also, $H0(x,a) = c(x,a) > 0$ for all $(x,a) \in \mathsf{G}$, which implies there exists an $\varepsilon > 0$ satisfying $H0 \geqslant \varepsilon(\|c\|/(1-\beta))\mathbb{1}$.

Example 6.2. Let $g\colon \mathbb{R} \to \mathbb{R}$ be an increasing and concave map such that $g(0) \geqslant 0$ and there is $K > 0$ satisfying $\|c\| + \|\beta\|g(K) < K$. Assume that c is positive

everywhere and bounded, and β is everywhere positive. Suppose that H is defined by

$$HQ(x,a) = c(x,a) + \mathbb{E}_{x,a}\beta(x,a,x')g\left(\min_{b\in\Gamma(x')} Q(x',b)\right) \qquad ((x,a)\in\mathsf{G}, Q\in\mathbb{R}^{\mathsf{G}}).$$

Then, we have $H(K\mathbb{1})(x,a) \leqslant \|c\| + \|\beta\|g(K) < K$ for any $(x,a)\in\mathsf{G}$. Moreover, we have $H0(x,a) \geqslant c(x,a) > 0$ for any $(x,a)\in\mathsf{G}$ so as there exists an $\varepsilon > 0$ satisfying $H0 \geqslant \varepsilon K\mathbb{1}$. This example allows discount factors to be greater than one. In this case, the convergence relies on the concavity of function g.

Lemma 6.1. *If Assumption 6.4 holds, then H defined by (20) satisfies Assumption 6.3, is globally stable on $[0, K\mathbb{1}]$, and has a unique fixed point $x^* \in [0, K\mathbb{1}]$.*

Lemma 6.2. *Suppose that Assumption 3.1 and 6.4 hold. If $\{Q_t\}$ is a sequence generated by (19) with $Q_0 \in [0, K\mathbb{1}]$, then $\{Q_t\}$ is in $[0, K\mathbb{1}]$ with probability 1.*

Proposition 6.1. *Let $\{Q_t\}$ be generated by (19). If Assumption 3.1 and 6.4 hold, then Q_t converges to Q^* with probability 1.*

Appendix A. Appendix

Proof of Lemma 2.1. Suppose $\rho(L_m) < 1$. Since $\rho(L_m) = \lim_{n\to\infty} \|L_m^n\mathbb{1}\|^{1/n}$, there exists n such that $\|L_m^n\mathbb{1}\| < 1$. Since $\beta(x,\sigma(x),x')P(x,\sigma(x),x') \leqslant \max_a \beta(x,a,x')P(x,a,x')$ for all $(x,x')\in\mathsf{X}^2$ and $\sigma\in\Sigma$, we have $d_1 \leqslant \sup_x L\mathbb{1}(x)$. Induction yields $d_n \leqslant \|L_m^n\mathbb{1}\| < 1$. □

Proof of Lemma 5.1. Assume that Assumption 3.3 holds and L is irreducible. Since L is irreducible and $\rho(L) < 1$, the Perron-Frobenius Theorem implies that there is a strictly positive eigenvector φ such that $L\varphi = \rho(L)\varphi$. That is, we have $\mathrm{diag}(\varphi)^{-1}L\varphi \leqslant \rho(L)\mathbb{1}$. We first show that T_σ is a contraction in $\|\cdot\|_\varphi$. Let $v, w \in \mathbb{R}^{\mathsf{X}}$ and $x\in\mathsf{X}$. By

definition, we have

$$\begin{aligned}
|T_\sigma v(x) - T_\sigma w(x)| &= \left| \sum_{x' \in \mathsf{X}} P(x, \sigma(x), x') \beta(x, \sigma(x), x') (v(x') - w(x')) \right| \\
&\leqslant \sum_{x' \in \mathsf{X}} P(x, \sigma(x), x') \beta(x, \sigma(x), x') |(v(x') - w(x')| \\
&\leqslant \sum_{x' \in \mathsf{X}} L(x, x') |(v(x') - w(x')| \\
&\leqslant \sum_{x' \in \mathsf{X}} L(x, x') \varphi(x') \max_{y \in \mathsf{X}} \frac{|(v(y) - w(y)|}{\varphi(y)} \\
&= \sum_{x' \in \mathsf{X}} L(x, x') \varphi(x') \|v - w\|_\varphi .
\end{aligned} \tag{21}$$

Dividing $\varphi(x)$ on both sides, we obtain

$$\frac{|T_\sigma v(x) - T_\sigma w(x)|}{\varphi(x)} \leqslant \frac{1}{\varphi(x)} \sum_{x' \in \mathsf{X}} L(x, x') \varphi(x') \|v - w\|_\varphi \leqslant \rho(L) \|v - w\|_\varphi ,$$

where the last inequality uses the fact $\mathrm{diag}(\varphi)^{-1} L \varphi \leqslant \rho(L) \mathbb{1}$. Taking the supremum on the left over X, we have $\|T_\sigma v - T_\sigma w\|_\varphi \leqslant \rho(L) \|v - w\|_\varphi$. Next, by rewriting (21), we have

$$T_\sigma v(x) \leqslant T_\sigma w(x) + \sum_{x' \in \mathsf{X}} L(x, x') \varphi(x') \|v - w\|_\varphi .$$

By taking the supremum over Σ of both sides, we obtain

$$Tv(x) \leqslant Tw(x) + \sum_{x' \in \mathsf{X}} L(x, x') \varphi(x') \|v - w\|_\varphi .$$

By interchanging the role of v and w and combining the two relations, we have

$$|Tv(x) - Tw(x)| \leqslant \sum_{x' \in \mathsf{X}} L(x, x') \varphi(x') \|v - w\|_\varphi .$$

The similar argument shows that $\|Tv - Tw\|_\varphi \leqslant \rho(L) \|v - w\|_\varphi$. Finally, we show that H is a contraction. Fix $Q, R \in \mathsf{X}^{\mathsf{G}}$. Then, we have

$$\begin{aligned}
|HQ(x, a) - HR(x, a)| &= \left| \sum_{x' \in \mathsf{X}} P(x, \sigma(x), x') \beta(x, \sigma(x), x') (\max_{a'} Q(x', a') - \max_{a'} R(x', a')) \right| \\
&\leqslant \sum_{x' \in \mathsf{X}} L(x, x') \max_{a'} |Q(x', a') - R(x', a')| \\
&\leqslant \sum_{x' \in \mathsf{X}} L(x, x') \varphi(x') \|Q - R\|_\varphi .
\end{aligned}$$

Again, the similar argument implies that $\|HQ - HR\|_\varphi \leqslant \rho(L)\|Q - R\|_\varphi$. □

Proof of Proposition 5.1. Suppose that the assumptions in the statement hold. Since Q^* is the fixed point of H, $r(x,a,x') \leqslant \bar{r}$ w.p.1, and $\mathrm{diag}(\varphi)^{-1}L\varphi = \rho(L)\mathbb{1}$, we have

$$\begin{aligned}
\left|\frac{Q^*(x,a)}{\varphi(x)}\right| &= \left|\sum_{x'} P(x,a,x')\left(\frac{r(x,a,x')}{\varphi(x)} + \frac{\beta(x,a,x')}{\varphi(x)}\max_b Q^*(x',b)\right)\right| \\
&\leqslant \left|\frac{\bar{r}}{\varphi(x)}\right| + \sum_{x'} P(x,a,x')\beta(x,a,x')\frac{\varphi(x')}{\varphi(x)}\max_b\left|\frac{Q^*(x',b)}{\varphi(x')}\right| \\
&\leqslant \|\bar{r}\|_\varphi + \sum_{x'} L(x,x')\frac{\varphi(x')}{\varphi(x)}\|Q^*\|_\varphi \\
&\leqslant \|\bar{r}\|_\varphi + \rho(L)\|Q^*\|_\varphi
\end{aligned}$$

for all $(x,a) \in \mathsf{G}$. Taking the maximum on the left, we have $\|Q^*\|_\varphi \leqslant \|\bar{r}\|_\varphi + \rho(L)\|Q^*\|_\varphi$, whence we obtain the first result. Next, with the assumption $\|Q_t\|_\varphi \leqslant \|\bar{r}\|_\varphi/(1-\rho(L))$, we obtain

$$\begin{aligned}
\left|\frac{\mathbb{E}(Q_{t+1}(x,a)|\mathscr{F}_t)}{\varphi(x)}\right| &= \Big|(1-\alpha_t)\frac{Q_t(x,a)}{\varphi(x)} \\
&\quad + \alpha_t\sum_{x'} P(x,a,x')\left(\frac{r(x,a,x')}{\varphi(x)} + \frac{\beta(x,a,x')}{\varphi(x)}\max_b Q_t(x',b)\right)\Big| \\
&\leqslant (1-\alpha_t)\|Q_t\|_\varphi + \alpha_t\left(\|\bar{r}\|_\varphi + \sum_{x'} L(x,x')\frac{\varphi(x')}{\varphi(x)}\|Q_t\|_\varphi\right) \\
&= (1-\alpha_t)\|Q_t\|_\varphi + \alpha_t(\|\bar{r}\|_\varphi + \rho(L)\|Q_t\|_\varphi) \\
&= \frac{\|\bar{r}\|_\varphi}{1-\rho(L)}
\end{aligned}$$

for all $(x,a) \in \mathsf{G}$. Taking the maximum on the left, we conclude that

$$\|\mathbb{E}(Q_{t+1}(\cdot,\cdot)|\mathscr{F}_t)\|_\varphi \leqslant \|\bar{r}\|_\varphi/(1-\rho(L)).$$

□

Let E be a real Banach space where a partial ordering is defined by a cone $P \subset E$ such that $x \leqslant y$ if and only if $y - x \in P$. We write $x < y$ if $x \leqslant y$ and $x \neq y$. A cone is called *normal* if there exists a constant $N > 0$ such that $\theta \leqslant x \leqslant y$ implies $\|x\| \leqslant N\|y\|$, where θ denotes the zero element of E. An operator $A\colon E \to E$ is called *increasing* if $x, y \in E$ with $x \leqslant y$ implies $Ax \leqslant Ay$. It is called *concave* if for any $x, y \in E$ with

$x \leqslant y$ and $\lambda \in [0,1]$, we have $A(\lambda x + (1-\lambda)y) \geqslant \lambda Ax + (1-\lambda)Ay$. For any $u_0, v_0 \in E$ with $u_0 < v_0$, we define an order interval by $[u_0, v_0] \coloneqq \{x \in E \colon u_0 \leqslant x \leqslant v_0\}$. Du (1990) shows the following fixed-point theorem with a concave operator (See also Zhang (2013) for the proof.)

Theorem A.1. *Suppose P is a normal cone, $u_0, v_0 \in E$, and $u_0 < v_0$. Moreover, $A \colon [u_0, v_0] \to E$ is an increasing operator. Let $h_0 = v_0 - u_0$. If A is a concave operator, $Au_0 \geqslant u_0 + \varepsilon h_0$, $Av_0 \leqslant v_0$ where $\varepsilon \in (0,1)$, then A has a unique fixed point $x^* \in [u_0, v_0]$. Moreover, for any $x_0 \in [u_0, v_0]$, the iterative sequence $\{x_n\}$ given by $x_n = Ax_{n+1}$ for $n \in \mathbb{N}$ satisfying that*

$$\|x_n - x^*\| \leqslant M(1-\varepsilon)^n \qquad (n \in \mathbb{N})$$

where $M = N^2\|h_0\| + (N+1)N\|B\theta\|\varepsilon^{-2}$, $\varepsilon \in (0,1)$ satisfies $B\theta = Au_0 - u_0 \geqslant \varepsilon h_0$, and N is the normal constant of P.

Proof of Theorem 6.1. Suppose that all the stated assumptions hold. Let $\{x_t\}$ be generated by (17). Define $U^{k+1} = HU^k$ and $L^{k+1} = HL^k$ for all $k \geqslant 0$ recursively with $U^0 = v$, $L^0 = u$. Assumption 6.3 implies $U^1 = Hv \leqslant v = U^0$ and $L^1 = Hu \geqslant u = L^0$. Since H is increasing, induction yields $L^k \leqslant L^{k+1}$ and $U^{k+1} \leqslant U^k$ for all k. Since the Du's Theorem A.1 implies that H is globally stable on $[u, v]$, we have $U^k \to x^*$ and $L^k \to x^*$. The conclusion then follows from the proof for Theorem 2 of Tsitsiklis (1994) that for every k, there exists some $t_k \in \mathbb{N}$ such that

$$L^k \leqslant x_t \leqslant U^k \qquad \text{for all } t \geqslant t_k. \tag{22}$$

□

Proof of Lemma 6.1. Let Assumption 6.4 hold. Then, H is a selfmap on $[0, K\mathbb{1}]$ and $H0 \geqslant \varepsilon K\mathbb{1}$. Let $\lambda \in [0,1]$ and fix $q_1, q_2 \in [0, K\mathbb{1}]$ with $q_1 \leqslant q_2$. Since φ is concave,

we have

$$\begin{aligned}
&H(\lambda q_1 + (1-\lambda)q_2)(x,a) \\
&\quad = \mathbb{E}_{x,a}\varphi\left(c(x,a,X') + \beta(x,a,X')\min_b\{\lambda q_1(X',b) + (1-\lambda)q_2(X',b)\}\right) \\
&\quad \geqslant \mathbb{E}_{x,a}\varphi\left(c(x,a,X') + \beta(x,a,X')\left(\lambda\min_b q_1(X',b) + (1-\lambda)\min_b q_2(X',b)\right)\right) \\
&\quad \geqslant \lambda\mathbb{E}_{x,a}\varphi\left(c(x,a,X') + \beta(x,a,X')\min_b q_1(X',b)\right) \\
&\qquad + (1-\lambda)\mathbb{E}_{x,a}\varphi\left(c(x,a,X') + \beta(x,a,X')\min_b q_2(X',b)\right) \\
&\quad = \lambda Hq_1(x,a) + (1-\lambda)Hq_2(x,a)
\end{aligned}$$

for all $(x,a) \in \mathsf{G}$. Therefore, H is concave. Since φ is increasing, H is also increasing. It follows from the Du's theorem A.1 that H has a unique fixed point Q^* in $[0, K\mathbb{1}]$ and there exists $\alpha \in (0,1)$ and $M > 0$ such that

$$\|H^m Q_0 - Q^*\| \leqslant \alpha^n M$$

for any $Q_0 \in [0, K\mathbb{1}]$ and $m \in \mathbb{N}$. □

Proof of Lemma 6.2. Let Assumption 3.1 and 6.4 hold. Fix $Q_0 \in [0, t\mathbb{1}]$. Suppose that Q_t is in $[0, K\mathbb{1}]$ for some t. Then, induction hypothesis and (19) imply

$$\begin{aligned}
Q_{t+1}(x,a) &= (1-\alpha_t(x,a))Q_t(x,a) + \alpha_t(x,a)\varphi\left(c_t + \beta_t\min_b Q(x',b)\right) \\
&\geqslant (1-\alpha_t(x,a))0 + \alpha_t(x,a)\varphi(c_t) \geqslant 0
\end{aligned}$$

and

$$\begin{aligned}
Q_{t+1}(x,a) &= (1-\alpha_t(x,a))Q_t(x,a) + \alpha_t(x,a)\varphi\left(c_t + \beta_t\min_b Q(x',b)\right) \\
&\leqslant (1-\alpha_t(x,a))K + \alpha_t(x,a)\varphi(c_t + \beta_t K) \\
&\leqslant (1-\alpha_t(x,a))K + \alpha_t(x,a)\varphi(\|c\| + \|\beta\|K) \\
&\leqslant (1-\alpha_t(x,a))K + \alpha_t(x,a)K = K
\end{aligned}$$

for all $(x,a) \in \mathsf{G}$. Therefore, we conclude $Q_t \in [0, K\mathbb{1}]$ for all $t \geqslant 0$ with probability 1. □

Proof of Proposition 6.1. Suppose that Assumption 3.1 and 6.4 hold and let $\{Q_t\}$ be generated by (19). We first rewrite Q_{t+1} by

$$Q_{t+1}(x_t,a_t) = (1-\alpha_t)Q_t(x_t,a_t) + \alpha_t\left(HQ(x_t,a_t) + w_t(x_t,a_t)\right)$$

where operator H is defined by (20) and w_t is defined by

$$w_t(x_t, a_t) = \varphi\left(c_t + \beta_t \min_{b\in\Gamma(x_{t+1})} Q_t(x_{t+1}, b)\right) - \mathbb{E}_{x_t,a_t}\varphi\left(c_t + \beta_t \min_{b\in\Gamma(x_{t+1})} Q_t(x_{t+1}, b)\right).$$

Clearly, we have $\mathbb{E}[w_t|\mathscr{F}_t] = 0$, where $\mathscr{F}_t$ is defined by (7). Since φ is concave, there are constants $p, q > 0$ such that $\varphi(s) \leqslant p + qs$ for all s. Hence, we have $\varphi^2(s) \leqslant (p + qs)^2$. Since in addition $\mathrm{Var}(\beta_t|\mathscr{F}_t)$ and $\mathrm{Var}(c_t|\mathscr{F}_t)$ are finite, there are constants A and B such that $\mathbb{E}[w_t^2|\mathscr{F}_t] \leqslant A + B\|Q_t\|^2$. Therefore, Assumption 6.2 holds. The conclusion then follows from Theorem 6.1, Lemma 6.1, and Lemma 6.2. □

References

Bertsekas, D. (2022): *Abstract Dynamic Programming*, Athena Scientific.

Bertsekas, D. P. and J. N. Tsitsiklis (1995): "Neuro-dynamic Programming: an Overview," in *Proceedings of 1995 34th IEEE Conference on Decision and Control*, IEEE, vol. 1, 560–564.

Calvano, E., G. Calzolari, V. Denicolo, and S. Pastorello (2020): "Artificial Intelligence, Algorithmic Pricing, and Collusion," *American Economic Review*, 110, 3267–3297.

Campbell, J. Y. and J. Ammer (1993): "What Moves the Stock and Bond Markets? A Variance Decomposition for Long-term Asset Returns," *The Journal of Finance*, 48, 3–37.

Charpentier, A., R. Elie, and C. Remlinger (2021): "Reinforcement Learning in Economics and Finance," *Computational Economics*, 1–38.

Cochrane, J. (2009): *Asset Pricing: Revised Edition*, Princeton University Press.

Cochrane, J. H. (2011): "Presidential Address: Discount Rates," *The Journal of Finance*, 66, 1047–1108.

Du, Y. (1990): "Fixed Points of Increasing Operators in Ordered Banach Spaces and Applications," *Applicable Analysis*, 38, 1–20.

Durdu, C. B., E. G. Mendoza, and M. E. Terrones (2009): "Precautionary Demand for Foreign Assets in Sudden Stop Economies: An Assessment of the New Mercantilism," *Journal of Development Economics*, 89, 194–209.

Hansen, L. P. and E. Renault (2010): "Pricing Kernels," *Encyclopedia of Quantitative Finance*.

Hasselt, H. (2010): "Double Q-learning," *Advances in Neural Information Processing Systems*, 23.

Hills, T. S. and T. Nakata (2018): "Fiscal Multipliers at the Zero Lower Bound: the Role of Policy Inertia," *Journal of Money, Credit and Banking*, 50, 155–172.

Hills, T. S., T. Nakata, and S. Schmidt (2019): "Effective Lower Bound Risk," *European Economic Review*, 120, 103321.

Hubmer, J., P. Krusell, and A. A. Smith Jr (2021): "Sources of US Wealth Inequality: Past, Present, and Future," *NBER Macroeconomics Annual*, 35, 391–455.

Jaakkola, T., M. Jordan, and S. Singh (1993): "Convergence of Stochastic Iterative Dynamic Programming Algorithms," *Advances in Neural Information Processing Systems*, 6.

Jasso-Fuentes, H., R. R. López-Martínez, and J. A. Minjárez-Sosa (2022): "Some Advances on Constrained Markov Decision Processes in Borel Spaces with Random State-dependent Discount Factors," *Optimization*, 1–27.

Lucas Jr, R. E. (1978): "Asset Prices in an Exchange Economy," *Econometrica: Journal of the Econometric Society*, 1429–1445.

Melo, F. S. (2001): "Convergence of Q-learning: A Simple Proof," *Institute Of Systems and Robotics, Tech. Rep*, 1–4.

Mendoza, E. G. (1991): "Real Business Cycles in a Small Open Economy," *The American Economic Review*, 797–818.

Minjárez-Sosa, J. A. (2015): "Markov Control Models with Unknown Random State-Action-Dependent Discount Factors," *Top*, 23, 743–772.

Nakata, T. (2016): "Optimal Fiscal and Monetary Policy with Occasionally Binding Zero Bound Constraints," *Journal of Economic Dynamics and Control*, 73, 220–240.

Neuneier, R. (1997): "Enhancing Q-learning for Optimal Asset Allocation," *Advances in Neural Information Processing Systems*, 10.

Obstfeld, M. (1990): "Intertemporal Dependence, Impatience, and Dynamics," *Journal of Monetary Economics*, 26, 45–75.

Park, D. and D. Ryu (2022): "Supply Chain Ethics and Transparency: An Agent-Based Model Approach with Q-learning Agents," *Managerial and Decision Economics*, 43, 3331–3337.

Rosenberg, J. V. and R. F. Engle (2002): "Empirical Pricing Kernels," *Journal of Financial Economics*, 64, 341–372.

Sargent, T. J. and J. Stachurski (2023): "Dynamic Programming Volume 1," QuantEcon, Available at https://dp.quantecon.org/ or https://github.com/QuantEcon/book-dp1.

Schmitt-Grohé, S. and M. Uribe (2003): "Closing Small Open Economy Models," *Journal of International Economics*, 61, 163–185.

Sharma, A., R. Gupta, K. Lakshmanan, and A. Gupta (2021): "Transition Based Discount Factor for Model Free Algorithms in Reinforcement Learning," *Symmetry*, 13, 1197.

Singh, S., T. Jaakkola, M. L. Littman, and C. Szepesvári (2000): "Convergence Results for Single-step On-policy Reinforcement-learning Algorithms," *Machine learning*, 38, 287–308.

Stachurski, J. and J. Zhang (2021): "Dynamic Programming with State-dependent Discounting," *Journal of Economic Theory*, 192, 105190.

Toda, A. A. (2021): "Perov's Contraction Principle and Dynamic Programming with Stochastic Discounting," *Operations Research Letters*, 49, 815–819.

——— (2023): "Unbounded Markov Dynamic Programming with Weighted Supremum Norm Perov Contractions," *ArXiv Preprint ArXiv:2310.04593*.

Tsitsiklis, J. N. (1994): "Asynchronous Stochastic Approximation and Q-learning," *Machine learning*, 16, 185–202.

Vasilev, A. (2022): "A Real-Business-Cycle Model with Endogenous Discounting and a Government Sector," *Notas Económicas*, 73–86.

Waltman, L. and U. Kaymak (2008): "Q-learning Agents in a Cournot Oligopoly Model," *Journal of Economic Dynamics and Control*, 32, 3275–3293.

Watkins, C. J. and P. Dayan (1992): "Q-learning," *Machine learning*, 8, 279–292.

Watkins, C. J. C. H. (1989): "Learning From Delayed Rewards," Ph.D. thesis, King's College, Cambridge United Kingdom.

Wei, Q. and X. Guo (2011): "Markov Decision Processes with State-dependent Discount Factors and Unbounded Rewards/Costs," *Operations Research Letters*, 39, 369–374.

Wu, X. and J. Zhang (2016): "Finite Approximation of the first Passage Models for Discrete-time Markov Decision Processes with Varying Discount Factors," *Discrete Event Dynamic Systems*, 26, 669–683.

Wu, X., X. Zou, and X. Guo (2015): "First Passage Markov Decision Processes with Constraints and Varying Discount Factors," *Frontiers of Mathematics in China*, 10, 1005–1023.

Yoshida, N., E. Uchibe, and K. Doya (2013): "Reinforcement Learning with State-dependent Discount Factor," in *2013 IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, IEEE, 1–6.

Zhang, Z. (2013): *Variational, Topological, and Partial Order Methods with Their Applications*, vol. 29 of *Developments in Mathematics*, Springer Berlin Heidelberg.