On Convergence of Average-Reward Off-Policy Control Algorithms in Weakly-Communicating MDPs

author names withheld

Under Review for OPT 2022

Abstract

We show two average-reward off-policy control algorithms, Differential Q Learning (Wan, Naik, & Sutton 2021a) and RVI Q Learning (Abounadi Bertsekas & Borkar 2001), converge in weakly-communicating MDPs. Weakly-communicating MDPs are the most general class of MDPs that a learning algorithm with a single stream of experience can guarantee obtaining a policy achieving optimal reward rate. The original convergence proofs of the two algorithms require that all optimal policies induce unichains, which is not necessarily true for weakly-communicating MDPs. To the best of our knowledge, our results are the first showing average-reward off-policy control algorithms converge in weakly-communicating MDPs. As a direct extension, we show that average-reward options algorithms introduced by (Wan, Naik, & Sutton 2021b) converge if the Semi-MDP induced by options is weakly-communicating.

1. Introduction

We consider off-policy control algorithms for the average-reward objective. These algorithms learn a policy that achieves the best possible average-reward rate, using data generated by some other policy that the agent may not have control of. Designing convergent off-policy algorithms for the average-reward objective is challenging. While there are several off-policy learning algorithms in the literature (Schwartz 1993; Singh 1994; Bertsekas & Tsitsiklis 1996; Das 1999; Ren & Krogh 2001; Gosavi 2004; Yang 2016), the only known convergent algorithms are SSP Q-learning and RVI Q-learning, both by Abounadi, Bertsekas, & Borkar (2001), and Differential Q-learning by Wan, Naik, & Sutton (2021a).

The convergence of SSP Q-learning is limited in MDPs with a state being recurrent under all policies. The convergence of the RVI Q-learning algorithm (Abounadi et al. 2001) was developed for unichain MDPs, which just means that the Markov chain induced by any stationary policy is unichain [1]. The convergence of Differential Q-learning (Wan et al. 2021a) requires a weaker assumption – all optimal policies being unichain. It is clear that RVI Q-learning also converges under this assumption with a small modification of its proof. If an optimal policy induces multiple recurrent classes, the proofs of RVI Q-learning and Differential Q-learning would not go through. Technically, this is because both two proofs require that the uniqueness of the solution for the action-value function up to an additive constant in the average-reward optimality equation.

1. A Markov chain is unichain if there is only one recurrent class in Markov chain, plus a possibly empty set of transient states.

© .

A more general class of MDPs, called weakly-communicating MDPs, is not limited in the number of recurrent classes under the optimal policy. In these MDPs, all policies may induce multiple recurrent classes. The only assumption is that, except for a set of states that are transient under every policy, all states are reachable from every other state in a finite number of steps with a non-zero probability. It has been observed that the set of weakly-communicating MDPs is the most general set of MDPs such that there exists a learning algorithm that can, using a single stream of experience, guarantee to identify a policy that achieves the optimal average reward rate in the MDP (Barlett & Tewari 2009).

In this paper, we show convergence of RVI Q-learning and Differential Q-learning in weakly-communicating MDPs, without requiring any additional assumptions compared with their original convergence results. Two key steps in our proof are 1) showing that the solution sets of the two algorithms are non-empty, closed, bounded, and connected, and 2) showing that 0 is the unique solution for the Bellman optimality equation when all rewards are 0. With these two results, we use asynchronous stochastic approximation results by Borkar (2009) to show convergence to the solution sets. As a direct extension of the above results, we also show the convergence of two algorithms that extend the Differential Q-learning algorithm to the options framework, introduced by Wan et al. (2021b), if the Semi-MDP induced by the MDP and the set of options is weakly-communicating.

2. Preliminaries

Consider a finite Markov decision process, defined by the tuple $\mathcal{M} \doteq (\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$, where $\mathcal{S}$ is a set of states, $\mathcal{A}$ is a set of actions, $\mathcal{R}$ is a set of rewards, and $p : \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the dynamics of the environment. Every time steps t the agent observes the state of the MDP $S_t \in \mathcal{S}$ and chooses an action $A_t \in \mathcal{A}$ using some policy $b : \mathcal{A} \times \mathcal{S} \to [0, 1]$, then receives from the environment a reward $R_{t+1} \in \mathcal{R}$ and the next state $S_{t+1} \in \mathcal{S}$, and so on. The transition dynamics are defined as $p(s', r \mid s, a) \doteq \Pr(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a)$ for all $s, s' \in \mathcal{S}, a \in \mathcal{A}$, and $r \in \mathcal{R}$. Denote the set of stationary Markov policies Π.

The reward rate of a policy π starting from a given start state s can be defined as:

$$r(\pi, s) \doteq \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}[R_t \mid S_0 = s, A_{0:t-1} \sim \pi]. \tag{1}$$

We say a set of states *communicating*, if there exists a policy such that moving from either one state in the set to the other one in the set in a finite number of steps has positive probability. If the entire state space is communicating, we say an MDP *communicating*. *Weakly-communicating* MDPs generalizes over communicating MDPs. In weakly-communicating MDPs, in addition to a *closed* communicating set of states, there is a possibly empty set of states that are under every policy.

For weakly-communicating MDPs, there exists a unique optimal reward rate r_*, which does not depend on the start state. We say a policy is optimal if it achieves r_* regardless of the start state. The goal of an off-policy control algorithm is to learn an optimal policy from the stream of experience $\ldots, S_t, A_t, R_{t+1}, S_{t+1}, \ldots$ generated by a behavior policy that is not necessarily the same as the agent's learned policy. Both RVI Q-learning and Differential Q-learning achieve this

goal by solving $\bar{r}$ and q in the optimality equation:

$$q(s,a) = \sum_{s',r} p(s',r \mid s,a)(r - \bar{r} + \max_{a'} q(s',a')), \quad \forall\, s \in \mathcal{S}, a \in \mathcal{A}. \tag{2}$$

It is known that r_* is the unique solution of $\bar{r}$ and any greedy policy w.r.t. any solution of q is an optimal policy. In addition, shifting any solution of q by any constant vector results in the other solution of q. Finally, unlike in unichain MDPs (or MDPs where all optimal policies are unichain) where all solutions are different by some constant vector, in weakly-communicating MDPs, solutions of q may have multiple degrees of freedom. That is, if q_1, q_2 are both solutions of q, it is possible that $q_1 \neq q_2 + ce, \forall\, c \in \mathbb{R}$, where e denotes the all-one vector.

If the agent has a set of options, it may choose to execute these options. Each option o in $\mathcal{O}$ has two components: the option's *policy* $\pi^o : \mathcal{A} \times \mathcal{S} \to [0,1]$, and the termination probability $\beta^o : \mathcal{S} \to [0,1]$. For simplicity, for any $s \in \mathcal{S}, o \in \mathcal{O}$, we use $\pi(a \mid s,o)$ to denote $\pi^o(a,s)$ and $\beta(s,o)$ to denote $\beta^o(s)$. If the agent executes option o at state s, the option's policy is followed, until the option terminates. Let $\mathcal{L}$ be the set of all possible lengths of options and $\hat{\mathcal{R}}$ be the set of all possible cumulative rewards. Note that $\mathcal{L}$ and $\hat{\mathcal{R}}$ are possibly countably infinite. Let $\hat{p}(s',r,l \mid s,o)$ be, when executing option o starting from state s, the probability of terminating at state s', with cumulative reward r and length l. Formally, for any $s, s' \in \mathcal{S}, o \in \mathcal{O}, r \in \hat{\mathcal{R}}, l \in \mathcal{L}$, $\hat{p}$ can be defined recursively in the following way:

$$\begin{aligned}\hat{p}(s',r,l \mid s,o) &= \sum_a \pi(a \mid s,o) \sum_{\tilde{s},\tilde{r}} p(\tilde{s},\tilde{r} \mid s,a) \\ &[\beta(\tilde{s},o)\mathbf{I}(\tilde{s} = s', \tilde{r} = r, \tilde{l} = 1) + (1 - \beta(\tilde{s},o))\hat{p}(s', r - \tilde{r}, l - 1 \mid \tilde{s},o)],\end{aligned} \tag{3}$$

where $\mathbf{I}$ is an indicator function.

An MDP $\mathcal{M}$ and a set of options $\mathcal{O}$ results in an Semi-MDP (SMDP) $\hat{\mathcal{M}} = (\mathcal{S}, \mathcal{O}, \mathcal{L}, \hat{\mathcal{R}}, \hat{p})$.

If the agent chooses options using a meta-policy (policy-over-options) $\pi : \mathcal{O} \times \mathcal{S} \to [0,1]$ and execute these options, we denote $\ldots, \hat{S}_n, \hat{O}_n, \hat{R}_{n+1}, \hat{S}_{n+1}, \ldots$ as the sequence of option transitions. For this SMDP, the *reward rate* of π given a start state s can be defined as $r^C(\pi,s) \doteq \lim_{t\to\infty} \mathbb{E}_\pi[\sum_{i=1}^t R_i \mid S_0 = s]/t$. $r(\pi,s) \doteq \lim_{n\to\infty} \mathbb{E}_\pi[\sum_{i=0}^n \hat{R}_i \mid \hat{S}_0 = s]/\mathbb{E}_\pi[\sum_{i=0}^n \hat{L}_i \mid \hat{S}_0 = s]$. Both limits exist and are equivalent (Puterman's (1994) propositions 11.4.1) under the following assumption:

Assumption 1 *For each option $o \in \mathcal{O}$, when executing the option, there is a non-zero probability of terminating the option after at most $|\mathcal{S}|$ stages, regardless of the initial states.*

Proposition 1 *Under Assumption 1, the expected value as well as the variance of the execution time and cumulative reward of every option at every state exist and are finite.*

We say an SMDP is weakly-communicating if the MDP with state space $\mathcal{S}$, action space $\mathcal{O}$, reward space $\hat{\mathcal{R}}$, and transition function $\sum_l \hat{p}(s,r,l \mid s,o)$ is weakly-communicating. Just like the MDP setting, if the SMDP is weakly-communicating, the optimal reward rate $\hat{r}_* \doteq \sup_{\pi\in\hat{\Pi}} r(\pi,s)$, where $\hat{\Pi}$ denotes the set of stationary Markov meta-policies, does not depend on the start state s. In addition, the solutions of q may not be different for a constant. Given an MDP and a set of options, the goal of the off-policy control problem is to find a policy that achieves $\hat{r}_*$. Inter-option

Differential Q-learning achieves this goal by solving the *optimality* equation for SMDPs (Puterman 1994):

$$q(s,o) = \sum_{s',r,l} \hat{p}(s',r,\, l \mid s,o)\big(r - \bar{r}\cdot l + \max_{o'} q(s',o')\big), \tag{4}$$

where q and $\bar{r}$ denote estimates of the option-value function and the reward rate respectively. Just like the MDP setting, $\bar{r}$ has $\hat{r}_*$ as its unique solution and solutions of q may not be different by a constant vector.

Intra-option Differential-learning finds an optimal policy by solving the *intra-option optimality* equation.

$$q(s,o) = \sum_{a} \pi(a \mid s,o) \sum_{s',r} p(s',r \mid s,a)(r - \bar{r} + u_*^q(s',o)), \quad \forall\, s \in \mathcal{S}, o \in \mathcal{O}, \tag{5}$$

where

$$u_*^q(s',o) \doteq \big(1 - \beta(s',o)\big)q(s',o) + \beta(s',o)\max_{o'} q(s',o'). \tag{6}$$

The following proposition shows that the set of solutions of (5) is the same as that of (4).

Proposition 2 *Any solution of* (4) *is also a solution of* (5) *and vice versa.*

3. Convergence Results

In this section, we present the convergence theory of Differential Q-Learning, RVI Q-Learning in weakly-communicating MDPs and that of the two option extensions of Differential Q-Learning in weakly-communicating SMDPs. Empirical results verifying the convergence of the two MDP algorithms are presented in Appendix L.

Differential Q-learning updates a table of estimates $Q_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as follows:

$$\begin{aligned} Q_{t+1}(S_t,A_t) &\doteq Q_t(S_t,A_t) + \alpha_{\nu(t,S_t,A_t)}\delta_t, \\ Q_{t+1}(s,a) &\doteq Q_t(s,a),\ \forall\, s,a \neq S_t,A_t, \end{aligned} \tag{7}$$

where $\nu(t,S_t,A_t)$ is the number of times S_t, A_t has been visited before time step t, $\alpha_{\nu(t,S_t,A_t)}$ is a step-size sequence, and δ_t, the temporal-difference (TD) error, is:

$$\delta_t \doteq R_{t+1} - \bar{R}_t + \max_a Q_t(S_{t+1},a) - Q_t(S_t,A_t), \tag{8}$$

where $\bar{R}_t$ is a scalar estimate of r_*, updated by:

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta\alpha_{\nu(t,S_t,A_t)}\delta_t, \tag{9}$$

and η is a positive constant.

We now present the convergence theorem of Differential Q-learning. We first state the required assumptions, which are also required by the original convergence proof of Differential Q-Learning by Wan et al. (2021a).

Assumption 2 *For all $n \geq 0$, $\alpha_n > 0$, $\sum_{n=0}^{\infty} \alpha_n = \infty$, and $\sum_{n=0}^{\infty} \alpha_n^2 < \infty$.*

Assumption 3 *Let $[\cdot]$ denote the integer part of $(\cdot)$, for $x \in (0,1)$, $\sup_n \frac{\alpha_{[xn]}}{\alpha_n} < \infty$ and $\frac{\sum_{n=0}^{[ym]} \alpha_n}{\sum_{n=0}^{m} \alpha_n} \to 1$ uniformly in $y \in [x, 1]$.*

Assumption 4 *There exists $\Delta > 0$ such that*

$$\liminf_{n \to \infty} \frac{\nu(n, s, a)}{n+1} \geq \Delta,$$

a.s., for all $s \in \mathcal{S}, a \in \mathcal{A}$. Furthermore, for all $x > 0$, let $N(n, x) = \min\left\{m > n : \sum_{k=n+1}^{m} \alpha_k \geq x\right\}$, the limit $\lim_{n\to\infty} \left(\sum_{k=\nu(n,s,a)}^{\nu(N(n,x),s,a)} \alpha_k\right) / \left(\sum_{k=\nu(n,s',a')}^{\nu(N(n,x),s',a')} \alpha_k\right)$ exists a.s. for all $s, s' \in \mathcal{S}, a, a' \in \mathcal{A}$.

Theorem 1 *If $\mathcal{M}$ is communicating and Assumptions 1–4 hold, the Differential Q-learning algorithm (Equations 7–9) converges, almost surely, $\bar{R}_t$ to r_*, Q_t to the set of solutions of (2) and*

$$r_* - \bar{R}_0 = \eta \left(\sum_{s,a} q(s,a) - \sum_{s,a} Q_0(s,a) \right), \tag{10}$$

and $r(\pi_t, s)$ to r_*, for all $s \in \mathcal{S}$, where π_t is any greedy policy w.r.t. Q_t.

Remark: If the MDP is weakly-communicating, that is, it contains transient states, the agent eventually reaches the closed communicating state and never returns to the transient states. Elements in Q_n that are associated with the closed communicating set converge to a set that depends on the values of Q and $\bar{R}$ when the MDP reaches the closed communicating set for the first time. Other elements in Q_n would only be visited for a finite number of times and can not be guaranteed to converge to their correct values by any learning algorithms. Other conclusions of the theorem remain unchanged. This observation on weakly-communicating MDPs apply also to Theorems 2–4.

The update rules of RVI Q-Learning are

$$\begin{aligned} Q_{t+1}(S_t, A_t) &\doteq Q_t(S_t, A_t) + \alpha_{\nu(t,S_t,A_t)} \delta_t(S_t, A_t) \\ Q_{t+1}(s, a) &\doteq Q_t(s, a), \ \forall\, s, a \neq S_t, A_t, \end{aligned} \tag{11}$$

where

$$\delta_t(S_t, A_t) \doteq R_{t+1} - f(Q_t) + \max_a Q_t(S_{t+1}, a) - Q_t(S_t, A_t). \tag{12}$$

and $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ satisfying the following assumption.

Assumption 5 *1) f is L-Lipschitz, 2) there exists a positive scalar u s.t. $f(e) = u$ and $f(x+ce) = f(x) + cu$, 3) $f(cx) = cf(x)$.*

Theorem 2 *If $\mathcal{M}$ is communicating and Assumptions 1–5 hold, then the RVI Q-Learning algorithm (Equations 11–12) converges, almost surely, $\bar{R}_t$ to r_*, Q_t to the set of solutions of (2) and*

$$r_* = f(q), \tag{13}$$

and $r(\pi_t, s)$ to r_*, for all $s \in \mathcal{S}$, where π_t is any greedy policy w.r.t. Q_t.

References

Abounadi, J., Bertsekas, D., & Borkar, V. S. (2001). Learning Algorithms for Markov Decision Processes with Average Cost. *SIAM Journal on Control and Optimization*.

Bertsekas, D. P. (2007). *Dynamic Programming and Optimal Control third edition, volume II*. Athena Scientific.

Borkar, V. S. (1998). Asynchronous Stochastic Approximations. *SIAM Journal on Control and Optimization*.

Borkar, V. S. (2009). *Stochastic Approximation: A Dynamical Systems Viewpoint*. Springer.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons.

Schweitzer, P. J. (1971). Iterative solution of the functional equations of undiscounted Markov renewal programming. *Journal of Mathematical Analysis and Applications*.

Schweitzer, P. J., & Federgruen, A. (1978). The Functional Equations of Undiscounted Markov Renewal Programming. *Mathematics of Operations Research*.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning. *Artificial Intelligence*.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction.* MIT Press.

Wan, Y., Naik, A., & Sutton, R. S. (2021a). Learning and Planning in Average-Reward Markov Decision Processes. *International Conference on Machine Learning*.

Wan, Y., Naik, A., & Sutton, R. S. (2021b). Average-Reward Learning and Planning with Options. *Conference on Neural Information Processing Systems*.

Appendix A. Proof of Proposition 1

Note that the execution time of each option $o \in \mathcal{O}$ is the return of executing this option's policy in an stochastic shortest path MDP (SSP-MDP, Bertsekas 2007) with state space $\mathcal{S} + \{\bot\}$ ($\bot$ is the terminal state), action space $\mathcal{A}$ and transition function $\hat{p}$ satisfying:

$$\hat{p}(s', 1 \mid s, a) \doteq (1 - \beta(s', o)) \sum_r p(s', r \mid s, a), \quad \forall\, s' \neq \bot, a \in \mathcal{A}$$
$$\hat{p}(\bot, 1 \mid s, a) \doteq \sum_{s'} \beta(s', o) \sum_r p(s', r \mid s, a), \quad \forall\, s \neq \bot, a \in \mathcal{A}$$
$$\hat{p}(\bot, 0 \mid \bot, a) \doteq 1, \quad \forall\, a \in \mathcal{A}.$$

Also, note that by Assumption 1, the option's policy is a 'proper' policy (Bertsekas 2007) in the SSP-MDP. That it, when using the policy, the MDP reaches the terminal state eventually regardless of the start state. Because the expected value of every proper policy of an SSP MDP exists and is finite (Section 2.1 of Bertsekas (2007)), the expected value of the execution time of option o exists. The existence of the variance can be shown using similar arguments as those used to show the existence of the expectation in Section 2.1 of Bertsekas (2007).

Similarly, the cumulative reward of each option o is the return of executing o's policy in a SSP-MDP with state space $\mathcal{S} + \{\bot\}$, action space $\mathcal{A}$, and transition function $\hat{p}$ satisfying:

$$\hat{p}(s', r \mid s, a) \doteq (1 - \beta(s', o)) p(s', r \mid s, a), \quad \forall\, s' \neq \bot, a \in \mathcal{A}$$
$$\hat{p}(\bot, r \mid s, a) \doteq \sum_{s'} \beta(s', o) p(s', r \mid s, a), \quad \forall\, s \neq \bot, a \in \mathcal{A}$$
$$\hat{p}(\bot, 0 \mid \bot, a) \doteq 1, \quad \forall\, a \in \mathcal{A}.$$

Again the option's policy is proper and expected value of the cumulative reward of option o exists. So does the variance of the cumulative reward.

Appendix B. Proof of Proposition 2

By the definition of $\hat{p}$ of the SMDP induced by choosing options in an MDP,

$$\hat{p}(\tilde{s}, \tilde{r}, \tilde{l} \mid s, o) = \sum_a \pi(a \mid s, o) \sum_{s', r} p(s', r \mid s, a)$$
$$[\beta(s', o)\mathbf{I}(s' = \tilde{s}, \tilde{r} = r, \tilde{l} = 1) + (1 - \beta(s', o))\hat{p}(\tilde{s}, \tilde{r} - r, \tilde{l} - 1 \mid s', o)]$$

$$
\begin{aligned}
q(s,o) &= \sum_{s',r,l} \hat{p}(s',r,l \mid s,o)(r - l\bar{r} + \max_{o'} q(s',o')) \\
&= \sum_{\tilde{s},\tilde{r},\tilde{l}} \sum_{a} \pi(a \mid s,o) \sum_{s',r} p(s',r \mid s,a) \\
&\quad [\beta(s',o)\mathbf{I}(s' = \bar{s}, \tilde{r} = r, \tilde{l} = 1) + (1-\beta(s',o))\hat{p}(\tilde{s},\tilde{r}-r,\tilde{l}-1 \mid s',o)](\tilde{r} - \tilde{l}\bar{r} + \max_{o'} q(s',o')) \\
&= \sum_{a} \pi(a \mid s,o) \sum_{s',r} p(s',r \mid s,a)\beta(s',o)(r - \bar{r} + \max_{o'} q(s',o')) \\
&\quad + \sum_{a} \pi(a \mid s,o) \sum_{s',r} p(s',r \mid s,a)(1-\beta(s',o)) \sum_{\tilde{s},\tilde{r},\tilde{l}} \hat{p}(\tilde{s},\tilde{r}-r,\tilde{l}-1 \mid s',o)](\tilde{r} - \tilde{l}\bar{r} + \max_{o'} q(s',o')) \\
&= \sum_{a} \pi(a \mid s,o) \sum_{s',r} p(s',r \mid s,a)\beta(s',o)(r - \bar{r} + \max_{o'} q(s',o')) \\
&\quad + \sum_{a} \pi(a \mid s,o) \sum_{s',r} p(s',r \mid s,a)(1-\beta(s',o)) \sum_{\tilde{s},\tilde{r},\tilde{l}} \hat{p}(\tilde{s},\tilde{r},\tilde{l} \mid s',o)](\tilde{r} + r - (\tilde{l}+1)\bar{r} + \max_{o'} q(s',o')) \\
&= \sum_{a} \pi(a \mid s,o) \sum_{s',r} p(s',r \mid s,a)\beta(s',o)(r - \bar{r} + \max_{o'} q(s',o')) \\
&\quad + \sum_{a} \pi(a \mid s,o) \sum_{s',r} p(s',r \mid s,a)(1-\beta(s',o)) \left(r - \bar{r} + \sum_{\tilde{s},\tilde{r},\tilde{l}} \hat{p}(\tilde{s},\tilde{r},\tilde{l} \mid s',o)](\tilde{r} - \tilde{l}\bar{r} + \max_{o'} q(s',o')) \right) \\
&= \sum_{a} \pi(a \mid s,o) \sum_{s',r} p(s',r \mid s,a)\beta(s',o)(r - \bar{r} + \max_{o'} q(s',o')) \\
&\quad + \sum_{a} \pi(a \mid s,o) \sum_{s',r} p(s',r \mid s,a)(1-\beta(s',o)) \big(r - \bar{r} + q(s',o)\big) \\
&= \sum_{a} \pi(a \mid s,o) \sum_{s',r} p(s',r \mid s,a)(r - \bar{r} + u_*^q(s',o))
\end{aligned}
$$

Therefore any solution of (4) must be a solution of (5)-(6) and vice versa.

Appendix C. SMDP Convergence Results

Consider option extensions of Differential Q-Learning. Given an SMDP $\hat{\mathcal{M}} = (\mathcal{S}, \mathcal{O}, \mathcal{L}, \hat{\mathcal{R}}, \hat{p})$, inter-option Differential Q learning maintains estimates of option values, and, inspired by Schweitzer (1971), update estimates using scaled TD errors:

$$
Q_{n+1}(\hat{S}_n, \hat{O}_n) \doteq Q_n(\hat{S}_n, \hat{O}_n) + \alpha_{\nu(n,\hat{S}_n,\hat{O}_n)} \delta_n / L_n(\hat{S}_n, \hat{O}_n), \tag{14}
$$

$$
Q_{n+1}(s,o) \doteq Q_n(s,o), \ \forall\, s, o \neq \hat{S}_n, \hat{O}_n,
$$

$$
\bar{R}_{n+1} \doteq \bar{R}_n + \eta \alpha_{\nu(n,\hat{S}_n,\hat{O}_n)} \delta_n / L_n(\hat{S}_n, \hat{O}_n), \tag{15}
$$

where $\nu(n, \hat{S}_n, \hat{O}_n)$ is the number of visits before stage n, $L_n(\cdot,\cdot)$ comes from an additional vector of estimates $L : \mathcal{S} \times \mathcal{O} \to \mathbb{R}$ that approximates the expected lengths of state-option pairs, updated

from experience by:

$$L_{n+1}(\hat{S}_n, \hat{O}_n) \doteq L_n(\hat{S}_n, \hat{O}_n) + \beta_{\nu(n,\hat{S}_n,\hat{O}_n)}(\hat{L}_n - L_n(\hat{S}_n, \hat{O}_n)), \tag{16}$$

where $\{\beta_n\}$ is an another step-size sequence. The TD-error δ_n in (14) and (15) is

$$\delta_n \doteq \hat{R}_n - L_n(\hat{S}_n, \hat{O}_n)\bar{R}_n + \max_o Q_n(\hat{S}_{n+1}, o) - Q_n(\hat{S}_n, \hat{O}_n), \tag{17}$$

Theorem 3 *If $\hat{\mathcal{M}}$ is communicating, Assumptions 1–4 hold, except for using $\nu(n, s, o)$ instead of $\nu(t, s, a)$, and that $0 \leq \beta_n \leq 1$, $\sum_n \beta_n = \infty$, and $\sum_n \beta_n^2 < \infty$, inter-option Differential Q-learning (Equations 14-17) converges, almost surely, Q_n the set of solutions of (4) and*

$$\hat{r}_* - \bar{R}_0 = \eta(\sum_{s,o} q(s, o) - \sum_{s,o} Q_0(s, o)), \tag{18}$$

$\bar{R}_n$ to $\hat{r}_*$, and $r(\pi_n, s)$ to $\hat{r}_*$ where π_n is a greedy policy w.r.t. Q_n.

Intra-option Differential Q learning also maintain estimates of option-values. However, instead of updating the estimates using option transitions, it updates for all options using each action transition.

$$Q_{t+1}(S_t, o) \doteq Q_t(S_t, o) + \alpha_{\nu(t,S_t,o)}\rho_t(o)\delta_t(o), \quad \forall\, o \in \mathcal{O}, \tag{19}$$

$$Q_{t+1}(s, o) \doteq Q_t(s, o), \quad \forall\, s \in \mathcal{S}, o \in \mathcal{O}, \tag{20}$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \sum_{o \in \mathcal{O}} \alpha_{\nu(t,S_t,o)}\rho_t(o)\delta_t(o), \tag{21}$$

where α_t is a step-size sequence, $\rho_t(o) \doteq \frac{\pi(A_t|S_t,o)}{\pi(A_t|S_t,O_t)}$ is the importance sampling ratio, and:

$$\delta_t(o) \doteq R_{t+1} - \bar{R}_t + u_*^{Q_t}(S_{t+1}, o) - Q_t(S_t, o), \tag{22}$$

where u_* is defined in (6).

Theorem 4 *If $\hat{\mathcal{M}}$ is communicating, Assumptions 1–4 hold, except for using $\nu(t, s, o)$ instead of $\nu(t, s, a)$, intra-option Differential Q-learning (Equations 19-22) converges, almost surely, Q_t to the set of solutions of (4) and (18), $\bar{R}_t$ to $\hat{r}_*$, and $r(\pi_t, s)$ to $\hat{r}_*$ where π_t is a greedy policy w.r.t. Q_t.*

Appendix D. Proof sketch of Theorem 1-Theorem 4

In this section, we sketch the proof of Theorem 1-Theorem 4. It has been shown that all four algorithms introduced above are special cases of the *General RVI Q algorithm* (Wan et al. 2021a,b). They also showed that General RVI Q converges under an assumption that is not satisfied for weakly-communicating MDPs/SMDPs. In order to show convergence for weakly-communicating MDPs/SMDPs, we replace this assumption with three weaker assumptions that are satisfied for these MDPs/SMDPs. All other assumptions are the same as those used by Wan et al. (2021a,b) and can be verified for all four algorithms using their arguments. We present General RVI Q and

prove its convergence with the three new assumptions in Appendix I. The next step of the proof would be verifying the three new assumptions when casting General RVI Q to each of the four algorithms. This should be straightforward given our Theorem 5 and Lemma 3. We delay this part to Appendix J. With the three assumptions hold we have the conclusion part of the convergence theorem of General RVI Q holds for each of the four algorithms. The convergence of the reward rates of greedy policies w.r.t. the action/option-values follows the convergence of these values and is shown in Appendix K.

Appendix E. Characterization of the Solution Set

In this section, we characterize the sets that the algorithms described in the above section converges to. This section plays a key role in showing their convergence.

We consider the set of solutions of the SMDP optimality (4) and

$$c = f(q), \tag{23}$$

where $f : \mathcal{S} \times \mathcal{O} \to \mathbb{R}$ satisfies Assumption 5, c is an arbitrary real. Denote this set as $\mathcal{Q}_\infty$. It is clear that (4) generalizes (2) and (23) generalizes (10), (13), and (18). And thus the characterization of $\mathcal{Q}_\infty$ applies to sets that action/option values in the aforementioned algorithms are claimed to converge to in Theorem 1–Theorem 4.

It is known that if the SMDP is weakly-communicating, (4) has $\hat{r}_*$ as its unique solution for $\bar{r}$. For q, it has been shown by Schweitzer & Federgruen (1978) (we will refer this work for multiple times and thus we use a shorthand 'S&F' for simplicity from now on) in their Theorem 4.2 that the set of solutions for q in (4) is closed, unbounded, connected, and possibly non-convex. The next theorem characterizes $\mathcal{Q}_\infty$.

Theorem 5 *If the SMDP is weakly-communicating and f 1) is Lipschitz, and 2) there exists some $u \neq 0$ such that $f(x + ce) = f(x) + cu, \forall\, c \in \mathbb{R}$, $\mathcal{Q}_\infty$ is non-empty, closed, bounded, connected, and possibly non-convex.*

Before presenting the proof, first note that our convergence proof does not rely on the convexity property and we delay the proof of non-convexity to Appendix H.

Proof

First, $\mathcal{Q}_\infty$ is non-empty. To see this point, note that for any solution of q in Equation (2) , q_*, $q_* + ce$ is also a solution for any $c \in \mathbb{R}$ and thus there must be a c such that Equation (23) holds because $f(x + ce) = f(x) + c$ for any x, c.

$\mathcal{Q}_\infty$ is closed because the set of solutions of q in Equation (4) is closed by S&F, the set of solutions of q in Equation (23) is closed because f is Lipschitz, and the intersection of two closed sets is closed.

Boundedness

We now show that $\mathcal{Q}_\infty$ is bounded. To this end, it is enough to show that the solution set of v in

$$v(s) = \max_o \sum_{s',r,l} \hat{p}(s',r,l \mid s,o)(r - l\hat{r}_* + v(s')), \forall\, s \in \mathcal{S} \tag{24}$$

$$c = f(\tilde{r} + \tilde{P}v), \tag{25}$$

where c is any constant, and

$$\tilde{r}(s,o) \doteq \sum_{s',r,l} \hat{p}(s',r,l \mid s,o)(r - l\hat{r}_*)$$

$$\tilde{P}(s,o,s') \doteq \sum_{r,l} \hat{p}(s',r,l \mid s,o)$$

is bounded. Denote this set as $\mathcal{V}_\infty$. Once we show $\mathcal{V}_\infty$ is bounded, $\mathcal{Q}_\infty$ is also bounded because each solution for q can be obtained from a solution for v with a linear operation:

$$q(s,o) = \sum_{s',r,l} \hat{p}(s',r,l \mid s,o)(r - l\bar{r} + v(s')), \quad \forall\, s \in \mathcal{S}, o \in \mathcal{O}. \tag{26}$$

In order to show boundedness, we need the following two lemmas, which slightly modify Theorem 4.1 and 5.1 by S&F. To this end, we need to first introduce some definitions.

For any $\pi \in \hat{\Pi}$, P_π denotes the $|\mathcal{S}| \times |\mathcal{S}|$ transition probability matrix under policy π. That is,

$$P_\pi(s,s') \doteq \sum_{o,r,l} \pi(o \mid s)\hat{p}(s',r,l \mid s,o). \tag{27}$$

Define P_π^∞ to be the *limiting matrix* of P_π, which is the Cesaro limit of the sequence $\{P_\pi^i\}_{i=1}^\infty$:

$$P_\pi^\infty \doteq \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} P_\pi^i. \tag{28}$$

Because $\mathcal{S}$ is finite, the Cesaro limit exists and P_π^∞ is a stochastic matrix (has row sums equal to 1). Let $l_\pi(s) \doteq \sum_{o,s',r,l} \pi(o \mid s)\hat{p}(s',r,l \mid s,o)l$. And let the fundamental matrix $Z_\pi \doteq (I - P_\pi^\infty + P_\pi^\infty)^{-1} = I + \lim_{\gamma\uparrow 1} \sum_{n=1}^\infty \gamma^n (P_\pi^n - P_\pi^\infty)$. Let $b(v,\pi)_i \doteq [r_\pi - l_\pi \hat{r}_* + P_\pi^\infty v - v]_i$ and $\hat{\Pi}_*$ denote the set of optimal meta-policies.

The following two lemmas are the similar to Theorem 4.1 (c) and Theorem 5.1 in S&F, except that 1) the '$\max$' operates over the set of all optimal policies, instead of the set of all deterministic optimal policies as in Theorem 4.1 (c), and 2) we consider the set of weakly-communicating SMDPs while S&F considers general multi-chain SMDPs. The proofs are also essentially the same. For completeness, we provide the proofs for these two lemmas in Sections F–G.

Lemma 1 *If the SMDP is weakly-communicating, v is a solution of Equation (24) if and only if*

$$v(s) = \max_{\pi \in \hat{\Pi}_*} [Z_\pi(r_\pi - l_\pi \hat{r}_*) + P_\pi^\infty v](s), \forall\, s \in \mathcal{S}, \tag{29}$$

where $\hat{\Pi}_*$ denotes the set of optimal meta-policies.

Define R_π as the set of recurrent states for P_π^∞ (i.e., $R_\pi \doteq \{s \mid P_\pi^\infty(s,s) > 0\}$). Let $R_* \doteq \{s \mid s \in R_\pi \text{ for some policy } \pi \in \hat{\Pi}_*\}$. S&F shows that there exists a policy $\pi \in \hat{\Pi}_*$ such that $R_\pi = R_*$. Let $n(\pi)$ be the number of recurrent classes for P_π^∞ and let $n_* \doteq n(\pi)$. Denote the set of recurrent classes of P_π as $\{R_{*\alpha} \mid 1, 2, \cdots, n_*\}$. The following lemma shows that the solution set of Equation (24) has n_* degrees of freedom.

Lemma 2 *If the SMDP is weakly-communicating, suppose v and $v + x$ are both solutions of Equation (24), then there exists n_* constants $y_1, y_2, \cdots, y_{n_*}$ such that*

$$x(s) = y_\alpha, i \in R_{*\alpha}, \quad \alpha = 1, \ldots, n_* \tag{30}$$

$$x(s) = \max_{\pi \in \hat{\Pi}_*} Z_\pi b(v,\pi)_s + \sum_{\beta=1}^{n_*} \left(\sum_{s' \in R_{*\beta}} P_\pi^\infty(s,s') \right) y_\beta, \quad s \in \mathcal{S} \backslash R_*, \tag{31}$$

$$y_\alpha \geq Z_\pi b(v,\pi)_s + \sum_{\beta=1}^{n_*} \left(\sum_{s' \in R_{*\beta}} P_\pi^\infty(s,s') \right) y_\beta, \quad \alpha = 1, \ldots, n_*, s \in R_{*\alpha}, \pi \in \hat{\Pi}_*. \tag{32}$$

For any $\beta \in \{1, 2, \cdots, n_*\}$, note that there must exist a policy $\pi(\beta)$ such that $R_{*\beta}$ is the only one recurrent class under $\pi(\beta)$. To see this point, note that the SMDP is weakly-communicating, and thus we can modify π to obtain a new policy such that all states except for those in $R_{*\beta}$ are transient. Such a policy satisfies our requirement.

Applying this observation and Lemma 2, for any $v \in \mathcal{V}$, we have for any given $\beta \in \{1, \ldots, n_*\}$,

$$\begin{aligned} y_\alpha &\geq \max_{s \in R_{*\alpha},} Z_{\pi(\beta)} b(v, \pi(\beta))_s + \sum_{s' \in R_{*\beta}} P_{\pi(\beta)}^\infty(s,s') y_\beta \\ &= \max_{s \in R_{*\alpha},} Z_{\pi(\beta)} b(v, \pi(\beta))_s + y_\beta, \quad \forall\, \alpha = 1, \cdots, n_*. \end{aligned}$$

The first term $\max_{s \in R_{*\alpha},} Z_{\pi(\beta)} b(v, \pi(\beta))_s$ is a constant given v and $\pi(\beta)$. Therefore we see that, for any other solution $v + x$ of Equation (24), if y_β is arbitrarily large then $y_\alpha, \forall\, \alpha = 1, \cdots, n_*$ should also be arbitrarily large. This would violate the Lipschitz assumption on f. To see this point, let

$$\tilde{f}(v) \doteq f(\tilde{r} + \tilde{P} v). \tag{33}$$

Let L be the Lipschitz constant of f. L is also the Lipschitz constant of $\tilde{f}$ because $\tilde{P}$ is a stochastic matrix and is thus a non-expansion. Fix $v_1 \in \mathcal{V}_\infty$ and choose a $\tilde{v}_2 \in \mathcal{V}_\infty$, denote $m = \|v_1 - \tilde{v}_2\|$. Choose a sufficiently large $d > 0$ such that $\kappa d > L \|v_1 + de - v_2\| = L \|v_1 + de - \tilde{v}_2 - de\| = Lm$, where $v_2 \doteq \tilde{v}_2 + de$. Because $\tilde{f}(v_1 + de) = \tilde{f}(v_1) + \kappa d$, $\left\| \tilde{f}(v_1 + de) - \tilde{f}(v_2) \right\| = \left\| f(v_1) + \kappa d - \tilde{f}(v_2) \right\| = \kappa d > L \|v_1 + de - v_2\|$, which violates the Lipschitz assumption. Because the choice of β is arbitrary, $\mathcal{V}_\infty$ is upper bounded.

In addition, because the choice of β is arbitrary, we have for any $\alpha \in \{1, \ldots, n_*\}$,

$$y_\alpha \geq \max_{\beta \in \{1, \ldots, n_*\}} \max_{s \in R_{*\alpha},} Z_{\pi(\beta)} b(v, \pi(\beta))_s + y_\beta$$

If y_α is chosen to be arbitrarily small then y_β should also be arbitrarily small for all $\beta = 1, \cdots, n_*$ but again this is not allowed due to Equation (25) for the same reason. Therefore $y_\alpha, \forall\ \alpha \in \{1, \ldots, n_*\}$ can not be arbitrarily small. Thus $\mathcal{V}_\infty$ is lower bounded. Combining the upper bound and lower bound, $\mathcal{V}_\infty$ is bounded. Therefore $\mathcal{Q}_\infty$ is also bounded.

Connectedness

We now show that $\mathcal{Q}_\infty$ is connected. To this end, again it is enough to show that $\mathcal{V}_\infty$ is connected.

Define a function that maps from any $v \in \mathcal{V}$ to a solution in $\mathcal{V}_\infty$. Specifically, fix c, let $g : \mathcal{V} \to \mathcal{V}_\infty$ with $g(v) = v + xe$, where x is the solution of $\tilde{f}(v + xe) = c$ and $\tilde{f}$ is defined in Equation (33). Note that x is unique given v because $\tilde{f}(v + xe) = \tilde{f}(v) + \kappa x$ and $x = (c - \tilde{f}(v))/\kappa$.

We now show that g is Lipschitz continuous. Consider any $v_1, v_2 \in \mathcal{V}$. Let x_1, x_2 satisfy $v_1 + x_1 e = g(v_1)$ and $v_2 + x_2 e = g(v_2)$ respectively. Again x_1, x_2 are unique given v_1, v_2. Note that

$$\begin{aligned}
&\tilde{f}(v_1 + x_1 e) - \tilde{f}(v_2 + x_1 e) \\
&= \tilde{f}(v_1 + x_1 e) - \tilde{f}(v_2 + x_2 e) + x_1 - x_2 \\
&= c - c + x_1 - x_2 \\
&= x_1 - x_2
\end{aligned}$$

$$\begin{aligned}
|x_1 - x_2| &= |\tilde{f}(v_1 + x_1 e) - \tilde{f}(v_2 + x_1 e)| \\
&\leq L \left\| v_1 + x_1 e - v_2 - x_1 e \right\| \\
&= L \left\| v_1 - v_2 \right\|
\end{aligned}$$

$$\|g(v_1) - g(v_2)\| = \|v_1 + x_1 e - v_2 - x_2 e\| \leq \|v_1 - v_2\| + \|x_1 e - x_2 e\| = (L|\mathcal{S}| + 1) \|v_1 - v_2\| .$$

Therefore g is Lipschitz continuous with Lipschitz constant $L|\mathcal{S}| + 1$.

Finally, because $\mathcal{V}$ is connected and the image of any continuous function on a connected set is connected, $g(\mathcal{V})$ is connected. Note that every point in $g(\mathcal{V})$ belongs to $\mathcal{V}_\infty$ by definition. Every point in $\mathcal{V}_\infty$ is also a point in $g(\mathcal{V})$. To see this point, pick any $x \in \mathcal{V}_\infty$, we can see that $x \in \mathcal{V}$ and that $g(x) = x$. Therefore $\mathcal{V}_\infty = g(\mathcal{V})$ is connected.

Given that $\mathcal{V}_\infty$ is connected, $\mathcal{Q}_\infty$ should also be connected because $\mathcal{Q}_\infty$ is a linear transformation of $\mathcal{V}_\infty$ (see Equation 26).

■

The other result we will need to use to show convergence of the four algorithms introduced in the previous section is the following one. With this result, the stability of the algorithms can be established using the result by Borkar and Meyn (2000) (see also, Section 3.2 by Borkar 2009).

Lemma 3 *If an SMDP is weakly-communicating and all rewards are* 0*,* 0 *is the only element in* $\mathcal{Q}_\infty$.

Proof Given a weakly-communicating SMDP, by Lemma 1, any solution of the state-value optimality equation Equation (24) satisfies

$$v(s) \geq \sum_{s' \in R_*} P^{\infty}_{\pi^*}(s, s')v(s'), \forall\, s \in R_*, \pi \in \hat{\Pi}_*$$

where R_* is defined right after Lemma 1. Also, because all rewards are 0, R_* is the closed communicating class and $\hat{\Pi}_* = \hat{\Pi}$ contains all stationary policies.

Denote v_* as a solution of Equation (24). Now pick an arbitrary policy $\pi \in \hat{\Pi}$. Consider every recurrent class C under P_π, we have $\forall\, s \in C$, $v_*(s) \geq \sum_{s' \in C} d^C_\pi(s')v_*(s')$, where d^C_π denotes the stationary distribution of π in the recurrent class C. Sum both sides over s weighted by d^C_π, $\sum_s d^C_\pi(s)v_*(s) \geq \sum_s d^C_\pi(s) \sum_{s' \in C} d^C_\pi(s')v_*(s') = \sum_{s' \in C} d^C_\pi(s')v_*(s')$. Because l.h.s. = r.h.s. we have $v_*(s) = \sum_{s' \in C} d^C_\pi(s')v_*(s')$. Therefore $v_*(s) = v_*(s'), \forall\, s, s' \in C$.

Now for any $s, s' \in R_*$, there must exists a $\pi \in \hat{\Pi}_* = \hat{\Pi}$ such that there is a path from s to s' and a path from s' to s, because s, s' are in the same communicating class. Therefore s, s' are in the same recurrent class under P_π. Thus we conclude that $v_*(s) = v_*(s')$. Therefore $\forall\, s, s' \in R_*$, $v_*(s) = v_*(s')$. The transient states values are uniquely determined by values of states in R_*. Thus the solution set of v in the state-value optimality equation Equation (24) is also unique up to some additive constant.

The solution set of the option-value optimality equation Equation (4) is also unique up to some additive constant. To see this point, let $v(s) = \max_o q(s, o)$, then Equation (4) transforms to the state-value optimality equation. Therefore the solution of q in Equation (4) satisfies that $\max_o q(\cdot, o)$ is unique up to an additive constant. Let q_* be a solution, $q_*(s, o_1) = q_*(s, o_2), \forall\, s \in \mathcal{S}, o_1, o_2 \in \mathcal{A}$ because $\forall\, s, \in \mathcal{S}, o \in \mathcal{O}$:

$$\begin{aligned} q_*(s, o) &= \sum_{s', r} p(s', r \mid s, o) \max_{o'} q_*(s', o') \\ &= \sum_{s', r} p(s', r \mid s, o) \max_{o'} q_*(s, o') \\ &= \max_{o'} q_*(s, o'). \end{aligned}$$

Together with Equation (23), the solution of q is uniquely determined. To see that 0 is the unique solution of q, we can just verify that 0 is a solution of q in Equation (2) and Equation (23). The lemma is proved. ■

Appendix F. Proof of Lemma 1

Part 1: Equation (24) $\implies$ Equation (29):

Choose any solution of Equation (24), v_*, and choose any $\pi \in \hat{\Pi}_*$. Using Theorem 3.1 (e) by S&F,

$$\begin{aligned} 0 &\geq r_\pi - l_\pi \hat{r}_* + P_\pi v_* - v_* \\ (I - P_\pi)v_* &\geq r_\pi - l_\pi \hat{r}_* \end{aligned}$$

(in weakly-communicating SMDPs, the "L" set in Theorem 3.1 (e) is $\mathcal{O}$).

Using the above inequality and Lemma 2.1 by S&F, we have

$$\begin{aligned}(I - P_\pi^\infty)v_* &\geq Z_\pi(r_\pi - l_\pi \hat{r}_*)\\ v_* &\geq Z_\pi(r_\pi - l_\pi \hat{r}_*) + P_\pi^\infty v_*\end{aligned}$$

Because π can be any element in $\hat{\Pi}_*$,

$$v_*(s) \geq \sup_{\pi \in \hat{\Pi}_*} [Z_\pi(r_\pi - l_\pi \hat{r}_*) + P_\pi^\infty v_*](s), \forall\, s \in \mathcal{S}. \tag{34}$$

We now show that there exists a $\pi \in \hat{\Pi}_*$ such that

$$v_* = Z_\pi(r_\pi - l_\pi \hat{r}_*) + P_\pi^\infty v_*.$$

This result, along with Equation (34) shows that

$$v_*(s) = \max_{\pi \in \hat{\Pi}_*} [Z_\pi(r_\pi - l_\pi \hat{r}_*) + P_\pi^\infty v_*](s), \forall\, s \in \mathcal{S}.$$

Define $\hat{\Pi}(v_*) \subseteq \hat{\Pi}$ to be the set of policies achieving all maxima in Equation (24) with $v = v_*$. That is, for any $\pi \in \hat{\Pi}(v_*)$,

$$(I - P_\pi)v_* = r_\pi - l_\pi \hat{r}_*.$$

Multiplying both sides by Z_π, we have

$$Z_\pi(I - P_\pi)v_* = Z_\pi(r_\pi - l_\pi \hat{r}_*).$$

Because $Z_\pi(I - P_\pi) = I - P_\pi$ by Equation 2.2 in S&F, we have

$$v_* = Z_\pi(r_\pi - l_\pi \hat{r}_*) + P_\pi v_*.$$

Part 2: Equation (29) $\implies$ Equation (24):

Conversely, if v satisfies Equation (29), define

$$\tilde{v}(s) \doteq \max_o \sum_{s',r,l} \hat{p}(s', r, l \mid s, o)(r - l\hat{r}_* + v(s')), \forall\, s \in \mathcal{S}. \tag{35}$$

For any $\pi \in \hat{\Pi}_*$,

$$\begin{aligned}\tilde{v}(s) &= \max_o \sum_{s',r,l} \hat{p}(s', r, l \mid s, o)(r - l\hat{r}_* + v(s'))\\ &\geq [r_\pi - l_\pi \hat{r}_* + P_\pi v](s)\\ &\geq [r_\pi - l_\pi \hat{r}_* + P_\pi(Z_\pi(r_\pi - l_\pi \hat{r}_*) + P_\pi^\infty v)](s)\\ &= [(I + P_\pi Z_\pi)(r_\pi - l_\pi \hat{r}_*) + P_\pi^\infty v](s)\\ &= [(Z_\pi + P_\pi^\infty)(r_\pi - l_\pi \hat{r}_*) + P_\pi^\infty v](s)\\ &= [Z_\pi(r_\pi - l_\pi \hat{r}_*) + P_\pi^\infty v](s).\\ \implies \tilde{v} &\geq v.\end{aligned}$$

Here the second inequality follows Equation (29). The second equality follows Equation 2.2 in S&F and the third equality follows Theorem 3.1 (a) in S&F.

Now let π be a deterministic policy achieving all $|\mathcal{S}|$ maxima in Equation (35). Then

$$v \leq \tilde{v} = r_\pi - l_\pi \hat{r}_* + P_\pi v$$
$$(I - P_\pi)v \leq r_\pi - l_\pi \hat{r}_* v$$

Multiply both sides by $P_\pi^\infty \geq 0$, we have

$$0 = P_\pi^\infty (I - P_\pi)v \leq P_\pi^\infty (r_\pi - l_\pi \hat{r}_* e).$$

Also note that the r.h.s. ≤ 0 because

$$P_\pi^\infty (r_\pi - l_\pi \hat{r}_*) \leq P_\pi^\infty (r_\pi - l_\pi r(\pi)) = 0.$$

Combining the above two inequalities, we have

$$P_\pi^\infty (r_\pi - l_\pi \hat{r}_*) = 0. \tag{36}$$

By part (a) of Theorem 3.1 in S&F, $\pi \in \hat{\Pi}_*$. Then we have

$$v \leq r_\pi - l_\pi \hat{r}_* + P_\pi v.$$

By Lemma 2.1 in S&F, we have

$$v \leq Z_\pi (r_\pi - l_\pi \hat{r}_*) + P_\pi^\infty v.$$

Inserting this inequality to Equation (35), we have

$$\begin{aligned}
\tilde{v} &= [r_\pi - l_\pi \hat{r}_* + P_\pi v] \\
&\leq (r_\pi - l_\pi \hat{r}_* + P_\pi [Z_\pi (r_\pi - l_\pi \hat{r}_*) + P_\pi^\infty v]) \\
&= [(I + P_\pi Z_\pi)(r_\pi - l_\pi \hat{r}_*) + P_\pi^\infty v] \\
&= [(Z_\pi + P_\pi^\infty)(r_\pi - l_\pi \hat{r}_*) + P_\pi^\infty v] \\
&= [Z_\pi (r_\pi - l_\pi \hat{r}_*) + P_\pi^\infty v] \\
&\leq \max_{\pi \in \hat{\Pi}_*} [Z_\pi (r_\pi - l_\pi \hat{r}_*) + P_\pi^\infty v] \\
&= v.
\end{aligned}$$

Here the third equation follows Equation 2.2 in S&F and the fourth equation follows Equation (36).

Combining $\tilde{v} \geq v$ and $\tilde{v} \leq v$, we have $\tilde{v} = v$ and therefore $\tilde{v}(s) = \max_o \sum_{s',r,l} \hat{p}(s', r, l \mid s, o)(r - l\hat{r}_* + v(s')), \quad \forall\, s \in \mathcal{S}$. And thus v is a solution of Equation (24).

Appendix G. Proof of Lemma 2

Let $\hat{\Pi}_{R_*} \doteq \{\pi \in \hat{\Pi}_* \mid R_\pi = R_*, n(\pi) = n_*\}$. Choose $\pi \in \hat{\Pi}_{R_*}$. Because $v, v + x$ are both solutions of Equation (24), using part (e)(2) of Theorem 3.1 by S&F, we have, $\forall\, s \in R_*$,

$$v(s) = [r_\pi - l_\pi \hat{r}_* + P_\pi v](s)$$
$$[v + x](s) = [r_\pi - l_\pi \hat{r}_* + P_\pi (v + x)](s).$$

Subtracting, we have $x(s) = [P_\pi x](s)$. Iteratively applying this equation, we have

$$x(s) = [P_\pi^\infty x](s) = \langle d_\pi^\alpha, x\rangle, \forall\, s \in R_{*\alpha},$$

where d_π^α is the unique stationary distribution of the α-th recurrent class of P_π. This proves Equation (30).

Because $v + x$ is a solution of Equation (24), by Lemma 1, $\forall\, s \in \mathcal{S}$,

$$\begin{aligned}
[v+x](s) &= \max_{\pi\in\hat{\Pi}_*}[Z_\pi(r_\pi - l_\pi\hat{r}_*) + P_\pi^\infty[v+x]](s)\\
x(s) &= \max_{\pi\in\hat{\Pi}_*}[Z_\pi(r_\pi - l_\pi\hat{r}_* + P_\pi^\infty v - v) + P_\pi^\infty x](s)\\
&= \max_{\pi\in\hat{\Pi}_*}[Z_\pi b(v,\pi) + P_\pi^\infty x](s)\\
&= \max_{\pi\in\hat{\Pi}_*}\left([Z_\pi b(v,\pi)](s) + \sum_{\beta=1}^{n_*}\left(\sum_{s'\in R_{*\beta}} P_\pi^\infty(s,s')\right) y_\beta\right).
\end{aligned}$$

Therefore Equation (31) and Equation (32) hold.

Appendix H. Proof of Non-Convexity

We now show that both $\mathcal{V}_\infty$ and $\mathcal{Q}_\infty$ are not necessarily convex. Consider the MDP shown in the left sub-figure of Figure 1. This is a communicating MDP. The optimal reward rate is 0.

Let $c = 3$ in Equation (25) and let $f(q) = \sum_{s,a} q(s,a)$. The solution set $\mathcal{V}_\infty = \{v \in \mathbb{R}^3 \mid v(1) \geq -2 + v(2); v(2) \geq v(3); v(3) = \max(v(1), -1 + v(2)); 2v(1) + 3v(2) + v(3) = 6\}$, which corresponds to the two connected yellow line segments in the right sub-figure in Figure 1. From the plot we see that $\mathcal{V}_\infty$ is not convex.

Let s and d denote *solid* and *dashed* respectively. Consider two solutions $q_1, q_2 \in \mathcal{Q}_\infty$:

$$\begin{aligned}
q_1(1,s) = 1, q_1(1,d) = -1,\\
q_1(2,s) = 1, q_1(2,d) = 1,\\
q_1(3,s) = 0, q_1(3,d) = 1
\end{aligned}$$

$$\begin{aligned}
q_2(1,s) = -1/6, q_2(1,d) = -1/6,\\
q_2(2,s) = 11/6, q_2(2,d) = 5/6,\\
q_2(3,s) = 5/6, q_2(3,d) = -1/6.
\end{aligned}$$

Consider the mid point $\bar{q} = 0.5q_1 + 0.5q_2$.

$$\begin{aligned}
\bar{q}(1,s) = 5/12, \bar{q}(1,d) = -7/12,\\
\bar{q}(2,s) = 17/12, \bar{q}(2,d) = 11/12,\\
\bar{q}(3,s) = 5/12, \bar{q}(3,d) = 5/12.
\end{aligned}$$

Note that

$$11/12 = \bar{q}(2,d) \neq \max(-1 + \bar{q}(3,s), \bar{q}(3,d)) = 5/12.$$

Therefore $\bar{q}$ does not satisfy Equation (2) and $\bar{q} \neq \mathcal{Q}_\infty$. Thus $\mathcal{Q}_\infty$ is not convex.

(*a*)

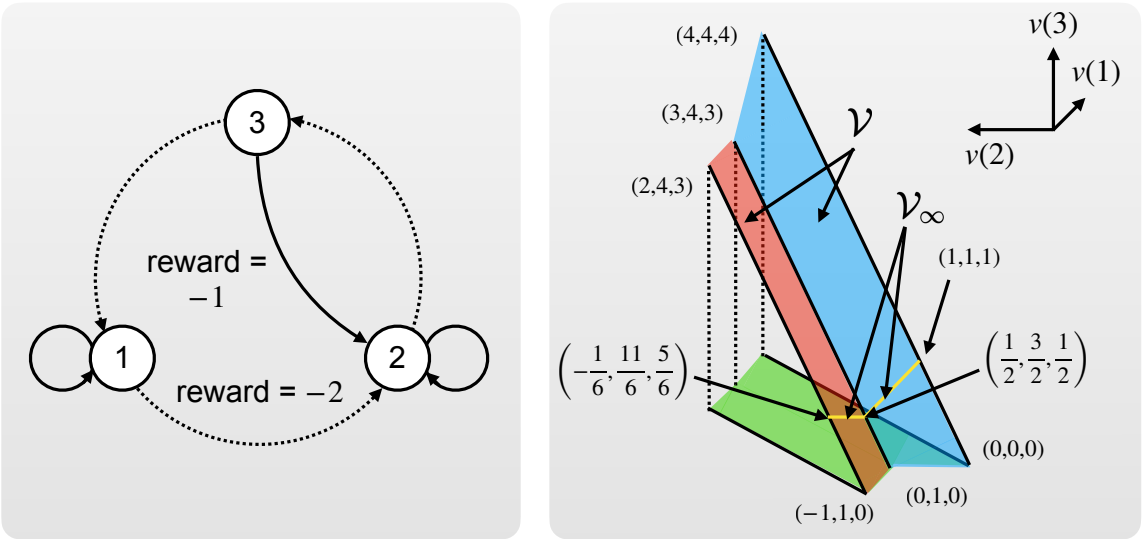

Figure 1: Illustration example. *Left*: the example MDP. There are three states marked by three circles respectively. There are two actions *solid* and *dashed*, both have deterministic effects. Taking action *solid* at state *3* results in a reward of -1. Taking action *dashed* at state *1* results in a reward of -2. All other rewards are 0. *Right*: a graphical explanation of $\mathcal{V}, \mathcal{V}_\infty$. The two yellow line segments together represent the solution set $\mathcal{V}_\infty$, which is not convex. The red and blue regions together represent $\mathcal{V}$.

Appendix I. General RVI Q

We now start to present General RVI Q.

Let $\mathcal{I} \doteq \{1, 2, \cdots, k\}$ where k is a positive integer. Consider solving $\bar{r} \in \mathbb{R}$ and $q \in \mathbb{R}^{|\mathcal{I}|}$ in following equation

$$r(i) - \bar{r} + g(q)(i) - q(i) = 0, \forall\, i \in \mathcal{I} \tag{37}$$

where $r \in \mathbb{R}^{|\mathcal{I}|}$ is any fixed $|\mathcal{I}|$-dim vector, and $g : \mathbb{R}^{|\mathcal{I}|} \to \mathbb{R}^{|\mathcal{I}|}$ satisfies Assumption 6.

We now consider an algorithm solving Equation (37). This algorithm maintains a $|\mathcal{I}|$-dim vector of estimates $Q \in \mathbb{R}^{|\mathcal{I}|}$, and updates Q using

$$Q_{n+1}(i) \doteq Q_n(i) + \alpha_{\nu(n,i)}\big(R_n(i) - F_n(Q_n) + G_n(Q_n)(i) - Q_n(i) + \epsilon_n(i)\big) I\{i \in Y_n\}, \tag{38}$$

where

1. $\{Y_n\}$ is the "update schedule" – it is a set-valued process taking values in the set of nonempty subsets of $\mathcal{I}$ with the interpretation: $Y_n = \{i : i^{\text{th}}$ component of Q was updated at time $n\}$,
2. $\nu(n, i) \doteq \sum_{k=0}^{n} \mathbb{I}\{i \in Y_k\}$, where $\mathbb{I}$ is the indicator function (i.e., $\nu(n, i) =$ the number of times the i component was updated up to step n)

3. $\{\alpha_n\}$ is a step-size sequence,
4. $\{R_n\}$ is a sequence of i.i.d. random vectors satisfying $\mathbb{E}\left[R_n\right] = r, \forall\, n = 0, 1, \ldots,$
5. for any $Q \in \mathbb{R}^{|\mathcal{I}|}$, $\{G_n(Q)\}$ is a sequence of i.i.d. random vectors satisfying $\mathbb{E}[G_n(Q)(i)] = g(Q)(i), \forall\, i \in \mathcal{I}, n = 0, 1, \ldots,$
6. for any $Q \in \mathbb{R}^{|\mathcal{I}|}$, $\{F_n(Q)\}$ is a sequence of i.i.d. random variables $\mathbb{E}[F_n(Q)] = f(Q), \forall\, n = 0, 1, \ldots$ where $f : \mathcal{I} \to \mathbb{R}$ is a function, and
7. ϵ_n is a sequence of random vectors of size $|\mathcal{I}|$.

We need the following assumptions in addition to Assumptions 1–5.

Assumption 6 *1) g is a max-norm non-expansion, 2) g is a span-norm non-expansion, 3) $g(x + ce) = g(x) + ce$ for any $c \in \mathbb{R}, x \in \mathbb{R}^{|\mathcal{I}|}$, 4) $g(cx) = cg(x)$ for any $c \in \mathbb{R}, x \in \mathbb{R}^{|\mathcal{I}|}$.*

Assumption 7 *For $n \in \{0, 1, 2, \ldots\}$, $\mathbb{E}[\|R_n - r\|^2] \leq K$, $\mathbb{E}[\|G_n(Q) - g(Q)\|^2] \leq K(1+\|Q\|^2)$ for any $Q \in \mathbb{R}^{|\mathcal{I}|}$, and $\mathbb{E}[\|F_n(Q) - f(Q)e\|^2] \leq K(1 + \|Q\|^2)$ for any $Q \in \mathbb{R}^{|\mathcal{I}|}$ for a suitable constant $K > 0$.*

We make following two assumptions on ϵ_n.

Assumption 8 $\mathbb{E}[\|\epsilon_n\|^2 \mid \mathcal{F}_n] < K(1 + \|Q_n\|^2)$ *a.s.. Further, ϵ_n converges to 0 a.s..*

We further make the following Assumptions 9 and 11 on the solution set of Equation (37).

Assumption 9 *Equation 37 has a unique solution for $\bar{r}$. That is, there exists a pair $\bar{r} \in \mathbb{R}, q \in \mathbb{R}^{|\mathcal{I}|}$ satisfying Equation (37) and if both $\bar{r}_1, q_1$ and $\bar{r}_2, q_2$ are solutions Equation (37), $\bar{r}_1 = \bar{r}_2$.*

Denoted the unique solution of $\bar{r}$ by $r_\#$.

Define $\mathcal{Q}_\#$ to be the set of $q \in \mathbb{R}^{|\mathcal{I}|}$ satisfying Equation (37) and

$$r_\# = f(q). \tag{39}$$

Assumption 10 *$\mathcal{Q}_\#$ is non-empty, bounded, closed, and connected.*

Assumption 11 *If $r(i) = 0, \forall\, i \in \mathcal{I}$, then 0 is the only element in $\mathcal{Q}_\#$.*

Assumption 1 is assumed in Section 2 of Wan et al. (2021b). Assumptions 2–8 are the same as Assumptions A.1-A.7 by Wan et al. (2021b). Assumptions 9–11 replace Assumption A.8 by Wan et al. (2021b).

Theorem 6 *Under Assumptions 1-11, General RVI Q Equation (38) converges, almost surely, Q_n to $\mathcal{Q}_\#$ and $f(Q_n)$ to $r_\#$.*

Proof The proof would by large degree repeats that of Theorem A.1 by Wan et al. (2021). For simplicity, we only highlight modifications.

Both our proof and the proof of Theorem A.1 study two ordinary differential equations (ODEs):

$$\dot{y}_t \doteq T_1(y_t) - y_t, \tag{40}$$

$$\dot{x}_t \doteq T_2(x_t) - x_t = T_1(x_t) - x_t + (r_\# - f(x_t))\, e, \tag{41}$$

where

$$\begin{aligned} T_1(Q)(i) &\doteq r(i) + g(Q)(i) - r_\#, \\ T_2(Q)(i) &\doteq r(i) + g(Q)(i) - f(Q) \\ &= T_1(Q)(i) + (r_\# - f(Q))\,. \end{aligned}$$

Lemma A.1, A.3 by Wan et al. (2021) and their proofs still hold (by replacing r_∞ with $r_\#$). Lemma A.2 by Wan et al. (2021) is replaced by the following one. The proof follows the same arguments with those for Lemma A.2.

Lemma 4 *The set of equilibrium points of* (41) *is* $\mathcal{Q}_\#$.

Lemma A.4 by Wan et al. (2021) is replaced with the following two lemmas.

Lemma 5 *If* $r(i) = 0, \forall\, i \in \mathcal{I}$, 0 *is the globally asymptotically stable equilibrium for* (41).

Proof We have assumed that 0 is the only element in $\mathcal{Q}_\#$ in Assumption 11 and 0 is thus the unique equilibrium of (41). The rest follows the proof of Lemma A.4 by Wan et al. (2021) by replacing q_∞ with 0. ■

Lemma 6 $\mathcal{Q}_\#$ *is a compact connected internally chain transitive invariant set for the ODE* (41). *Futhermore, any set that contains points not in* $\mathcal{Q}_\#$ *is not a compact connected internally chain transitive invariant set for the ODE* (41).

Proof According to Assumption 10 $\mathcal{Q}_\#$ is closed and bounded and is thus compact. Assumption 10 also assumes that $\mathcal{Q}_\#$ is connected. $\mathcal{Q}_\#$ is an internally chain transitive invariant set for the ODE (41) because every element in $\mathcal{Q}_\#$ is an equilibrium point of the ODE and $\mathcal{Q}_\#$ is connected.

If a set contains a point $x \in \mathbb{R}^{|\mathcal{I}|}$ that is not in $\mathcal{Q}_\#$, this set can not be internally chain transitive because x is "transient" and the trajectory of (41) can not be arbitrarily close to x at some arbitrarily large time step. ■

The last step is to show a synchronous version of (38)

$$Q_{n+1}(i) \doteq Q_n(i) + \alpha_{\nu(n,i)}\big(R_n(i) - F_n(Q_n) + G_n(Q_n)(i) - Q_n(i) + \epsilon_n(i)\big) \tag{42}$$

converges. This result, together with Assumptions 3 4, guarantees convergence of the asynchronous update 42 by applying results from Section 7.4 by Borkar (2009)

Lemma 7 *Equation 42 converges a.s. Q_n to $\mathcal{Q}_\#$ as $n \to \infty$.*

Proof The proof essentially follows that of Lemma A.5 with two changes. First, using Lemma 5 we can show that the ODE $\dot{x}_t = h_\infty(x_t) = g(x_t) - f(x_t)e - x_t$ has the origin as the unique globally asymptotically stable equilibrium. Second, the proof of Lemma A.5 uses Borkar's (2009) Theorem 2, which proves that Q_n converges to a (possibly sample path dependent) compact connected internally chain transitive invariant set of $\dot{x}_t = h(x_t)$ where

$$h(Q_n)(i) \doteq r(i) - f(Q_n) + g(Q_n)(i) - Q_n(i).$$

Lemma 6 and with Theorem 2 by Borkar (2009) together imply that Q_n must converge to $\mathcal{Q}_\#$. ■

■

Appendix J. Verifying Assumption 9–Assumption 11

When casting General RVI Q to RVI Q-learning, (37) becomes (2). It is then clear that Assumption 9 satisfies because the associated MDP is communicating, Assumption 10 holds because of Theorem 5 and Assumption 11 holds because of Lemma 3.

When casting General RVI Q to Differential Q-learning, (37) becomes (2) for an MDP with the same transition dynamics as the original one and all rewards being shifted by a constant that depends on initial action-value estimate Q_0 and reward rate estimate $\bar{R}_0$. It is clear that this new MDP, just like the original one, is communicating. Therefore Theorem 5 and Lemma 3 hold and thus Assumption 9–Assumption 11 hold.

When casting General RVI Q to inter-option Differential Q-learning, (37) becomes (4) for an SMDP with the same transition dynamics as the original one and the reward of each state-option pair being shifted in proportional to its expected duration. Again the resulting SMDP is communicating and therefore Theorem 5 and Lemma 3 hold and thus Assumption 9–Assumption 11 hold.

When casting General RVI Q to intra-option Differential Q-learning, (37) becomes (5), which is the same as (4) by Proposition 2 for the reward shifted SMDP introduced in the previous paragraph.

Appendix K. Convergence of Reward Rates of Greedy Policies

Lemma 8 *Assume that the SMDP is weakly-communicating, suppose Q_n converges to $\mathcal{Q}_\infty$ almost surely, let π_n be a greedy policy w.r.t. Q_n, $r(\pi_n, s) \to \hat{r}_*$ almost surely.*

Proof We will need the following lemma for the proof.

Lemma 9 *For any $m \in \{1, 2, 3, \ldots\}$, $a, b \in \mathbb{R}^m$, $b > 0$, and for any $p \in \mathbb{R}^m$ such that $p \geq 0$, $\sum_s p(s) = 1$,*

$$\frac{p^\top a}{p^\top b} \geq \min_s \frac{a(s)}{b(s)},$$
$$\frac{p^\top a}{p^\top b} \leq \max_s \frac{a(s)}{b(s)}.$$

Proof For any $s \in \{1, 2, 3, \ldots, m\}$

$$\frac{a(s)}{b(s)} \geq \min_{s'} \frac{a(s')}{b(s')}$$
$$a(s) \geq b(s) \min_{s'} \frac{a(s')}{b(s')}$$
$$p(s)a(s) \geq p(s)b(s) \min_{s'} \frac{a(s')}{b(s')}$$

Therefore

$$p^\top a \geq p^\top b \min_{s'} \frac{a(s')}{b(s')}$$

By our assumptions on b and p, $p^\top b > 0$, we have

$$\frac{p^\top a}{p^\top b} \geq \min_{s'} \frac{a(s')}{b(s')}.$$

$\frac{p^\top a}{p^\top b} \leq \max_s \frac{a(s)}{b(s)}$ can be shown in the same way. ■

Given that Q_n converges to $\mathcal{Q}_\infty$, consider $r(\pi_n)$ where π_n is a greedy policy w.r.t. Q_n. We show that $r(\pi_n)$ converges to $\hat{r}_*$.

For any $\pi \in \hat{\Pi}$, P_π denotes the $|\mathcal{S}| \times |\mathcal{O}| \times |\mathcal{S}| \times |\mathcal{O}|$ transition probability matrix under policy π. That is,

$$P_\pi(s, o, s', o') \doteq \sum_{r,l} \hat{p}(s', r, l \mid s, o)\pi(o' \mid s'). \tag{43}$$

Let P_π^∞ be the *limiting matrix* of P_π, which is the Cesaro limit of the sequence $\{P_\pi^i\}_{i=1}^\infty$:

$$P_\pi^\infty \doteq \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} P_\pi^i. \tag{44}$$

Because $\mathcal{S}$ is finite, the Cesaro limit exists and P_π^∞ is a stochastic matrix (has row sums equal to 1).

Let $r(s,o) \doteq \sum_{s',r,l} \hat{p}(s',r,l \mid s,o) r$ be the one-stage option reward and $l(s,o) \doteq \sum_{s',r,l} \hat{p}(s',r,l \mid s,o) l$ be the one-stage option length. By part (a) of Theorem 11.4.1 in Puterman (1994),

$$r(\pi_n, s, o) = \frac{P^{\infty}_{\pi_n} r(s,o)}{P^{\infty}_{\pi_n} l(s,o)}.$$

Thus we have,

$$\begin{aligned} r(\pi_n, s, o) &= \frac{P^{\infty}_{\pi_n} r(s,o)}{P^{\infty}_{\pi_n} l(s,o)} \\ &= \frac{P^{\infty}_{\pi_n}(r + P_{\pi_n} Q_n - Q_n)(s,o)}{P^{\infty}_{\pi_n} l(s,o)} \\ &\geq \min_{s',o'} \frac{(r + P_{\pi_n} Q_n - Q_n)(s',o')}{l(s',o')} \\ &= \min_{s',o'} \frac{(TQ_n - Q_n)(s',o')}{l(s',o')}, \end{aligned}$$

where the inequality holds because of Lemma 9, and $TQ_n(s,o) \doteq r(s,o) + \sum_{s',r,l} \hat{p}(s',r,l \mid s,o) \max_{o'} Q_n(s',o')$.

Let $r(\pi, s, o) \doteq \sum_{s',r,l} \hat{p}(s',r,l \mid s,o) r(\pi, s')$ be the reward rate of policy π starting from s, o. Because the SMDP is weakly-communicating, consider an deterministic optimal policy $\pi \in \hat{\Pi}_*$, $r(\pi, s) = \hat{r}_*, \forall\, s \in \mathcal{S}$. Therefore $r(\pi, s, o) = \hat{r}_*$. Now we have, for any s, o,

$$\begin{aligned} \hat{r}_* &= r(\pi, s, o) \\ &= \frac{P^{\infty}_{\pi^*} r(s,o)}{P^{\infty}_{\pi^*} l(s,o)} \\ &= \frac{P^{\infty}_{\pi^*}(r + P_{\pi} Q_n - Q_n)(s,o)}{P^{\infty}_{\pi^*} l(s,o)} \\ &\leq \max_{s',o'} \frac{(r + P_{\pi} Q_n - Q_n)(s',o')}{l(s',o')} \\ &\leq \max_{s',o'} \frac{(r + P_{\pi_n} Q_n - Q_n)(s',o')}{l(s',o')} \\ &= \max_{s',o'} \frac{(TQ_n - Q_n)(s',o')}{l(s',o')}. \end{aligned}$$

The first inequality holds because of Lemma 9 and the second inequality holds because π_n is a greedy policy w.r.t. Q_n.

With the above results, and that $\hat{r}_* \geq r(\pi_n, s, o)$, we have, for any s, o,

$$\min_{s',o'} \frac{(TQ_n - Q_n)(s',o')}{l(s',o')} \leq r(\pi_n, s, o) \leq \hat{r}_* \leq \max_{s',o'} \frac{(TQ_n - Q_n)(s',o')}{l(s',o')}$$

Therefore,

$$\max_{s,o} \hat{r}_* - r(\pi_n, s, o) \leq sp\left(\frac{TQ_n - Q_n}{l}\right),$$

where $sp(x) = \max_i x(i) - \min x(i)$ denotes the span of vector x.

Because $Q_n \to \mathcal{Q}_\infty$ a.s., every point q in $\mathcal{Q}_\infty$ satisfies $sp((Tq-q)/l) = 0$ because $(Tq-q)/l = \hat{r}_* e$ by (4). In addition, $sp((TQ_n - Q_n)/l)$ is a continuous function of Q_n, by continuous mapping theorem, $sp((TQ_n - Q_n)/l) \to 0$ a.s. Therefore we conclude that $r(\pi_n, s, o) \to \hat{r}_*$. By definition, $r(\pi_n, s) = \sum_o \pi_n(o \mid s) r(\pi_n, s, o)$. Therefore $r(\pi_n, s) \to \hat{r}_*$ ■

Appendix L. Empirical Results

In this section, we empirically verify our convergence results of Differential Q-learning and RVI Q-learning by showing the dynamics of the estimated action values of two algorithms in a communicating MDP and a weakly-communicating MDP.

We first consider the communicating MDP shown in Figure 2. For this MDP, we apply Differential Q-learning with initial action values 0, initial reward rate estimate -3, and $\eta = 1$. Then $c = \eta \sum_{s,a} Q(s,a) - \bar{R}_0 = 3$. The behavior chooses action *solid* with probability 0.8 and action *dashed* with probability 0.2 for both two states. The stepsize is 0.1. We performed 10 runs for each algorithm. Each run starts from state 1 and lasts for 1000 steps. Every 10 steps, we record the estimated action values and plot the higher action value of each state in the figure. Figure 3(left) shows the evolution of these action values. In the right panel of the same figure, we also show it using RVI Q-learning with action values being initialized with 0 and $f(q) = q(1, \textit{dashed})$. More detailed explanation is provided in the figure's caption. It can be seen that for both algorithms, 1) for each run, the estimated action values converges to a point in the solution set (the black line segments), 2) for different runs, the estimated action values generally converge to different points in the solution set.

We now consider apply both of thee two algorithms, with the same parameter settings and initialization in a weakly-communicating MDP (Figure 4), which is just the above communicating MDP plus a transient state. In the transient

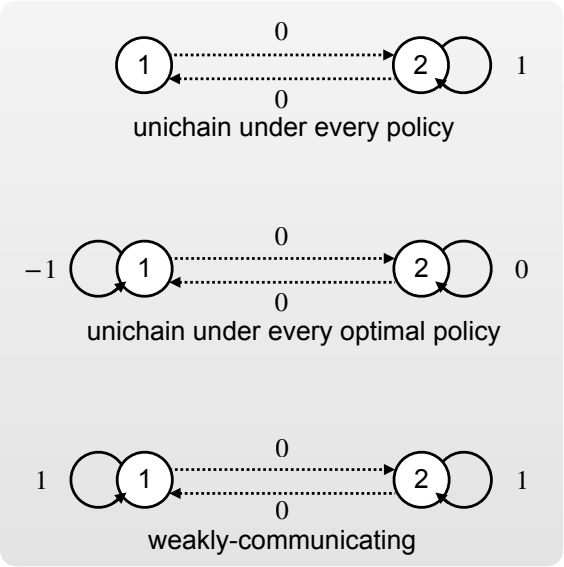

Figure 2: Examples of three different types of MDPs. In each of the three MDPs, there are two states marked by two circles respectively. There are two actions *solid* and *dashed*, both causing deterministic effects. *Top*: The MDP is unichain under every stationary policy. *Middle*: There are two deterministic optimal policies, moving to state 2 or going back and forth between state 1 and state 2. Other optimal stationary policies are mixtures of these two deterministic policies. All optimal policies are unichain. *Down*: There are three deterministic optimal policies: moving to state 1, moving to state 2, and staying at the current state. Other optimal stationary policies are mixtures of these two deterministic policies. The policy of staying at the current state induces two recurrent classes.

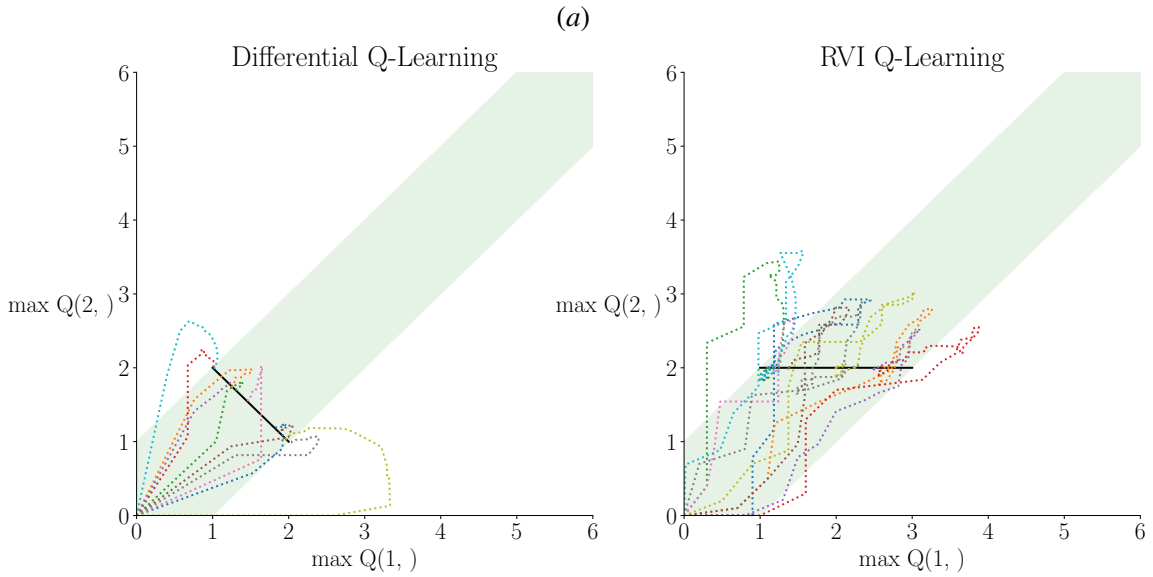

Figure 3: Evolution of estimated action values of Differential Q-Learning and RVI Q-Learning in the communicating MDP shown in the bottom of Figure 2. In each figure, the x and y axes are the higher estimated action value at state 1 and 2, respectively. The light green region marks the solution set of the optimality equation, the black line segment marks the solution set $\mathcal{Q}_\infty$. Each colored dotted trajectory marks the evolution of the estimated action values. starts from zero point and ends at some point on the black line segment.

state, taking both $solid$ and $dashed$ actions results in staying in the transient state with probability 0.9. The MDP moves to state 1 with probability 0.1 given action $solid$ and moves to state 2 with probability 0.1 given action $dashed$. The reward starting from state 0 is always -5. The agent starts interacting with the MDP from state 0. Because the agent could spends different amount of time in the transient state for different runs, the agent may enter in the communicating set, which contains state 1 and 2, with different action values associated with state 0.

The solution set of Differential Q-Learning depends on the action values associate with the transient states when entering the communicating class and therefore in the figure, the points the the action values converge to are not in a line. But action values in all runs converge to the region within the green region, which corresponds to the solution set of the optimality equation.

On the other hand, The solution set of RVI Q-learning with the choice of the reference function $f(q) = q(1, dashed)$ does not depend on the action values associate with the transient

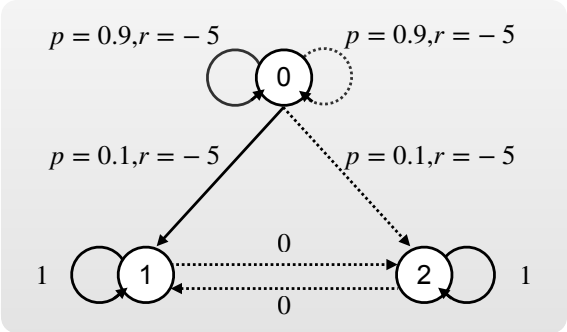

Figure 4: A weakly-communicating MDP modified from the communicating MDP shown at the bottom of Figure 2 by adding a transient state 0.

states when entering the communicating class.
Therefore the solution set remains unchanged.
Note that if we choose $f(q) = q(0, dashed)$, then again the solution set of RVI Q-learning has that dependence.

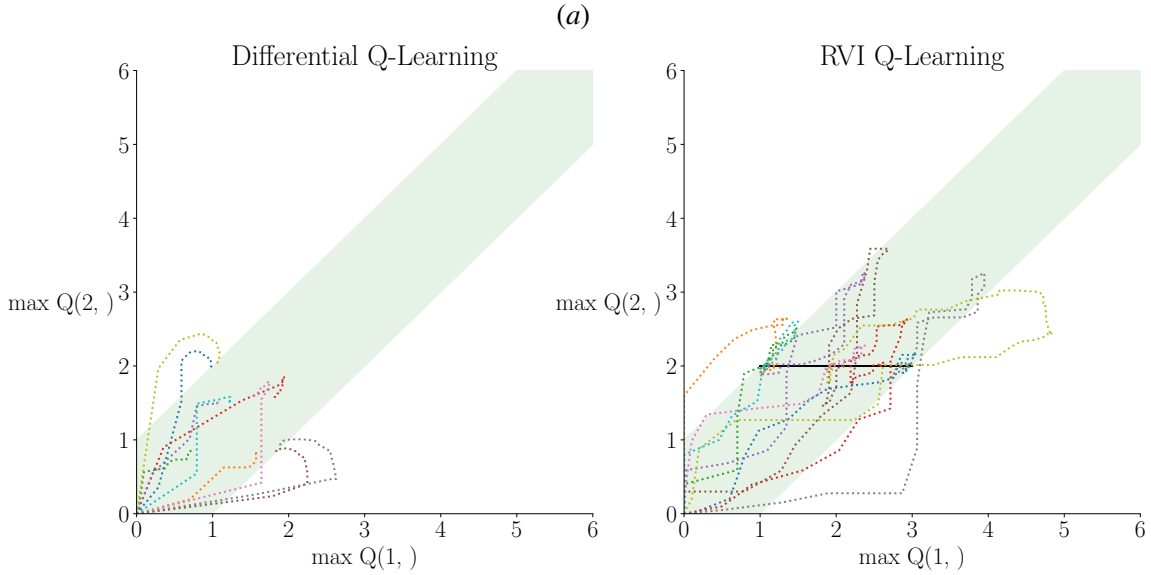

Figure 5: Evolution of estimated action values of Differential Q-Learning and RVI Q-Learning in the weakly-communicating MDP shown in the bottom of Figure 4.