OPT2022: 14th Annual Workshop on Optimization for Machine Learning

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

Yi Wan WAN6@UALBERTA.CA
University of Alberta
Richard S. Sutton RSUTTON@UALBERTA.CA
University of Alberta, DeepMind

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 MDPs that can be solved by a learning algorithm with a single stream of experience. The original convergence proofs of the two algorithms require that the solution set of the average-reward optimality equation only has one degree of freedom, 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 for temporal abstraction introduced by Wan, Naik, & Sutton (2021b) converge if the Semi-MDP induced by options is weakly communicating.

1. Introduction

In this paper, we extend the convergence results of two off-policy control algorithms for the average-reward objective from a sub-class of MDPs to the most general class of MDPs that could be solved by algorithms learning from a single stream of experience. 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, the only known convergent algorithms are SSP Q-learning and RVI Q-learning, both by Abounadi, Bertsekas, & Borkar (2001), the algorithm by Ren & Krogh (2001), and Differential Q-learning by Wan, Naik, & Sutton (2021a). Others either do not have a convergence theory (Schwartz 1993, Singh 1994; Bertsekas & Tsitsiklis 1996, Das 1999) or have incorrect proof (Yang 2016, Gosavi 2004). [1]

The algorithm by Ren & Krogh (2001) requires knowledge of properties of the MDP which are not typically known. The convergence of SSP Q-learning requires knowing a state that is 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 [2]. The convergence of Differential Q-learning (Wan et al. 2021a) requires a

1. See Appendix D in Wan et al. (2021a) for a discussion about Yang's proof and see Appendix D of this paper for a discussion about Gosavi's proof.
2. A Markov chain is unichain if there is only one recurrent class in the Markov chain, plus a possibly empty set of transient states.

© Y. Wan & R.S. Sutton.

weaker assumption – the solution set of the average-reward optimality equation (formally defined later in (2)) only has one degree of freedom (all the solutions are different by a constant vector). This assumption can be satisfied if, for example, all optimal policies are unichain. It is clear that RVI Q-learning also converges under this assumption. It is not rare that the solution set of the average-reward optimality equation has more than one degree of freedom (e.g., the MDP at the bottom of Figure 1). In this case, 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 of the action values up to an additive constant (one degree of freedom) in the average-reward optimality equation, so that there is a unique equalibrium in the ordinary differential equations associated with the two algorithms.

A more general class of MDPs, called weakly communicating MDPs, may have more than a single degree of freedom in the solution set of the associated optimality equation. The definition of these MDPs is also natural: except for a possibly empty 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).

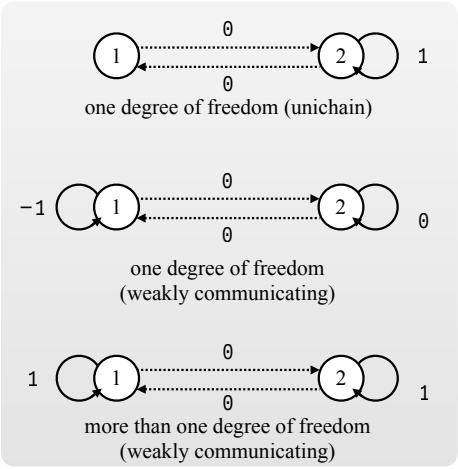

Figure 1: Examples of three different types of MDPs. In each of the three MDPs, there are two states marked by two circles respectively. There are two actions `solid` and `dashed`, both causing deterministic effects. *Top*: The solution set of q in the average-reward optimality equation ((2)) is $\{q(1,\texttt{dashed}) = c \quad 1, q(2,\texttt{solid}) = c, q(2,\texttt{dashed}) = c \quad 2, c \in \mathrm{R}\}$ and has one degree of freedom. The Markov chain under every stationary policy is unichain. *Middle*: The solution set of q is $\{q(1,\texttt{solid}) = c \quad 3, q(1,\texttt{dashed}) = c \quad 1, q(2,\texttt{solid}) = c, q(2,\texttt{dashed}) = c \quad 2, c \in \mathrm{R}\}$ and has one degree of freedom. The MDP is weakly communicating. *Down*: The solution set of q is has more than one degree of freedom. Note that both $\{q(1,\texttt{solid}) = 0, q(1,\texttt{dashed}) = 2, q(2,\texttt{solid}) = 1, q(2,\texttt{dashed}) = 1\}$ and $\{q(1,\texttt{solid}) = 0, q(1,\texttt{dashed}) = 1, q(2,\texttt{solid}) = 0, q(2,\texttt{dashed}) = 1\}$ are solutions of q in (2) and these two solutions are not different by a constant vector. The MDP is weakly communicating.

In this paper, we show the convergence of RVI Q-learning and Differential Q-learning in weakly communicating MDPs, without requiring any additional assumptions compared with their original convergence theories. 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 average-reward 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 a given MDP and a given set of options is weakly communicating.

2. Preliminaries

Consider a finite Markov decision process, defined by the tuple $M \doteq (S, A, R, p)$, where S is a set of states, A is a set of actions, R is a set of rewards, and $p : S \times R \times S \times A \to [0, 1]$ is the dynamics of the MDP. At each time step t, the agent observes the state of the MDP $S_t \in S$ and chooses an action $A_t \in A$ using some policy $b : A \times S \to [0, 1]$, then receives from the environment a reward $R_{t+1} \in R$ and the next state $S_{t+1} \in S$, 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 $s, s' \in S, a \in A$, and $r \in R$. 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:

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

Given an arbitrary MDP, the agent may not even be able to visit all states and would therefore miss the chance of learning, for every state s, a policy that achieves the optimal reward rate $\sup_\pi r(\pi, s)$ and the agent can at most learn an optimal policy for a set of states, each of which is reachable from every other state. Such a set of states is often called *communicating*. Formally speaking, 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 a positive probability. If the entire state space of an MDP is communicating, we say the MDP *communicating*. *Weakly communicating* MDPs generalize 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 transient 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 r and q in the optimality equation:

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

It is known that r^* is the unique solution of 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, where all solutions of q 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.

3. Convergence Results

In this section, we present convergence theories of Differential Q-learning and RVI Q-learning in weakly communicating MDPs, and theories 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 C.

Differential Q-learning updates a table of estimates $Q_t : S \quad A \; ! \; \mathbb{R}$ as follows:

$$\begin{aligned} Q_{t+1}(S_t; A_t) &\doteq Q_t(S_t; A_t) + \quad _{(t;S_t;A_t)} \; _t; \\ Q_{t+1}(s; a) &\doteq Q_t(s; a); \; 8 \, s; a \; 6= \; S_t; A_t; \end{aligned} \tag{3}$$

where $(t; S_t; A_t)$ is the number of times $S_t; A_t$ has been visited before time step t, $f \quad _{(t;S_t;A_t)}g$ is a step-size sequence, and $_t$, the temporal-difference (TD) error, is:

$$_t \doteq R_{t+1} \quad R_t + \max_a Q_t(S_{t+1}; a) \quad Q_t(S_t; A_t); \tag{4}$$

where R_t is a scalar estimate of r , updated by:

$$R_{t+1} \doteq R_t + \quad _{(t;S_t;A_t)} \; _t; \tag{5}$$

and $$ is a positive constant.

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

Assumption 1 *For all* $n \quad 0, \quad _n > 0, \; \sum_{n=0}^{1} \quad _n = 1$ *, and* $\sum_{n=0}^{1} \quad _n^2 < 1$ *.*

Assumption 2 *Let* $[\,]$ *denote the integer part of* $(\,)$*, for* $x \; 2 \; (0; 1)$*,* $\sup_n \frac{[xn]}{n} < 1$ *and* $\frac{\sum_{n=0}^{[ym]} \quad _n}{\sum_{n=0}^{m} \quad _n} \; ! \; 1$ *uniformly in* $y \; 2 \; [x; 1]$*.*

Assumption 3 *There exists* > 0 *such that*

$$\liminf_{n \; ! \; 1} \frac{(n; s; a)}{n + 1} \quad ;$$

a.s., for all $s \; 2 \; S; a \; 2 \; A$*. Furthermore, for all* $x > 0$*, let* $N(n; x) = \min \; m > n : \sum_{k=n+1}^{m} \quad _k \quad x \;$ *; the limit* $\lim_{n \; ! \; 1} \sum_{k=(n;s;a)}^{(N(n;x);s;a)} \quad _k = \sum_{k=(n;s^0;a^0)}^{(N(n;x);s^0;a^0)} \quad _k$ *exists a.s. for all* $s; s^0 \; 2 \; S; a; a^0 \; 2 \; A$*.*

Theorem 1 *If* M *is communicating and Assumptions 1–3 hold, the Differential Q-learning algorithm (Equations 3–5) converges, almost surely,* R_t *to* r *,* Q_t *to the set of solutions of* (2) *and*

$$r \quad R_0 = \sum_{s;a} q(s; a) \quad \sum_{s;a} Q_0(s; a) \; ! \; ; \tag{6}$$

and $r(\; _t; s)$ *to* r *, for all* $s \; 2 \; S$*, where* $_t$ *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 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 algorithm. Other conclusions of the theorem remain unchanged. This observation on weakly communicating MDPs also applies to Theorems 2–4.

The update rules of RVI Q-learning are

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

where

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

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

Assumption 4 *1) f is L-Lipschitz, 2) there exists a positive scalar u s.t. $f(\epsilon) = u$ and $f(x + c\epsilon) = f(x) + cu$, and 3) $f(cx) = cf(x), \forall c \geq 0$.*

Theorem 2 *If M is communicating and Assumptions 1–4 hold, then the RVI Q-learning algorithm (Equations 7–8) converges, almost surely, R_t to r^*, Q_t to the set of solutions of (2) and*

$$r^* = f(q^*), \tag{9}$$

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.

4. Conclusions

In this paper, we provide, for the first time, convergence results of off-policy average-reward control algorithms in weakly communicating MDPs, which are known to be the most general class of MDPs in which it is possible that a learning algorithm can guarantee to obtain an optimal policy. Specifically, we show two existing algorithms, RVI Q-learning and Differential Q-learning, converge in weakly communicating MDPs. As an extension, in Appendix A, we also showed two off-policy average-reward options learning algorithms converge if the SMDP induced by the options is weakly communicating.

Acknowledgements

The authors were generously supported by DeepMind, Amii, NSERC, and CIFAR. The authors wish to thank Huizhen Yu and Abhishek Naik for discussing several important related papers and discussing ideas to address the technical challenges in the proof. Computing resources were provided by Compute Canada.

References

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

Bartlett, P., & Tewari, A. (2009). REGAL: a regularization based algorithm for reinforcement learning in weakly communicating MDPs. *Uncertainty in Artificial Intelligence*.

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

Bertsekas, D. P., Tsitsiklis, J. N. (1996). *Neuro-dynamic Programming*. 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.

Das, T. K., Gosavi, A., Mahadevan, S. Marchalleck, N. (1999). Solving semi-Markov decision problems using average reward reinforcement learning. *Management Science*.

Gosavi, A. (2004). Reinforcement learning for long-run average cost. *European Journal of Operational Research*.

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

Ren, Z., Krogh, B. H. (2001). Adaptive control of Markov chains with average cost. *IEEE Transactions on Automatic Control*.

Schwartz, A. (1993). A reinforcement learning method for maximizing undiscounted rewards. In *Proceedings of the International Conference on Machine Learning*.

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*.

Singh, S. P. (1994). Reinforcement learning algorithms for average-payoff Markovian decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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. Learning with Options

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

$$\hat{p}(s', r, l \mid s, o) \doteq \sum_a \pi(a \mid s, o) \sum_{s', r} p(s', r \mid s, a)$$
$$[\gamma(s', o)\mathbb{I}(s' = s', r = r, t = 1) + (1 - \gamma(s', o))\hat{p}(s', r - r, l - 1 \mid s', o)], \tag{10}$$

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

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

Given an MDP and a set of options, if the agent chooses options using a meta policy, which is a policy that chooses from options, $\mu : O \times S \to [0,1]$ and executes these options, we denote the sequence of option transitions by $\ldots, \hat{S}_n, \hat{O}_n, \hat{R}_{n+1}, \hat{S}_{n+1}, \ldots$. For the associated SMDP, the *reward rate* of μ given a start state s can be defined as $r^C(\mu, s) \doteq \lim_{t \to \infty} \mathbb{E}[\sum_{i=1}^{t} R_i \mid S_0 = s]/t$ or $r(\mu, s) \doteq \lim_{n \to \infty} \mathbb{E}[\sum_{i=0}^{n} \hat{R}_i \mid \hat{S}_0 = s] / \mathbb{E}[\sum_{i=0}^{n} \hat{L}_i \mid \hat{S}_0 = s]$. Both limits exist and are equivalent (by Puterman's (1994) Propositions 11.4.1 and 11.4.7) under the following assumption:

Assumption 5 *For each option $o \in O$, when executing the option, there is a non-zero probability of terminating the option after at most $|S|$ stages, regardless of the state at which this option is initiated.*

Proposition 1 *Under Assumption 5, the expected value as well as the variance of the execution time and cumulative reward of every option at each state exist and are finite.*

We say an SMDP is weakly communicating if the MDP with state space S, action space O, reward space $\hat{R}$, and transition function $\sum_l \hat{p}(s', r, l \mid s, o)$ is weakly communicating. Just as in the MDP setting, if the SMDP is weakly communicating, the optimal reward rate $r^* \doteq \sup_{\mu \in \Pi} r(\mu, s)$, where Π 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 by a constant vector. Given an MDP and a set of options, the goal of the off-policy control problem is to find a policy that achieves r^*. Inter-option Differential Q-learning achieves this goal by solving the *optimality* equation for SMDPs (Puterman 1994):

$$q(s, o) = \sum_{s', r, l} \hat{p}(s', r, l \mid s, o) \left[r - r \cdot l + \max_{o'} q(s', o') \right], \tag{11}$$

where q and r denote estimates of the option-value function and the reward rate respectively. Just as in the MDP setting, 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.

$$q(s,o) = \sum_a \mu(a \mid s,o) \sum_{s',r} p(s',r \mid s,a)(r - \hat{r} + u^q(s',o)), \quad \forall s \in S, o \in O, \tag{12}$$

where

$$u^q(s',o) \doteq (1 - \beta(s',o))\, q(s',o) + \beta(s',o) \max_{o'} q(s',o'). \tag{13}$$

The following proposition shows that the set of solutions of (12) is the same as that of (11).

Proposition 2 *Any solution of* (11) *is also a solution of* (12) *and vice versa.*

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

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

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

$$R_{n+1} \doteq R_n + \alpha_{(n,\hat{S}_n,\hat{O}_n)} \delta_n L_n(\hat{S}_n, \hat{O}_n), \tag{15}$$

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

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

where $\{\alpha_n\}$ is the other step-size sequence. The TD-error δ_n in (14) and (15) is

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

Theorem 3 *If $\hat{M}$ is communicating, Assumptions 1–3 and 5 hold, except for using $\eta(n,s,o)$ instead of $\eta(t,s,a)$, and that $0 < \alpha_n < 1$, $\sum_n \alpha_n = \infty$, and $\sum_n \alpha_n^2 < \infty$, inter-option Differential Q-learning (Equations 14-17) converges, almost surely, Q_n to the set of solutions of* (11) *and*

$$\hat{r} \quad R_0 = \sum_{s,o} q(s,o) \left(\sum_{s,o} Q_0(s,o) \right); \tag{18}$$

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 maintains estimates of option values. However, instead of updating the estimates using option transitions, it updates for all options using each action transition $(S_t; O_t; A_t; R_{t+1}; S_{t+1})$.

$$Q_{t+1}(S_t; o) \doteq Q_t(S_t; o) + \quad _{(t;S_t;o)} \ _t(o) \ _t(o); \quad 8\, o \,2\, O; \tag{19}$$

$$Q_{t+1}(s; o) \doteq Q_t(s; o); \quad 8\, s \,2\, S; o \,2\, O;$$

$$R_{t+1} \doteq R_t + \sum_{o2O} \quad _{(t;S_t;o)} \ _t(o) \ _t(o); \tag{20}$$

where $f \ _t g$ is a step-size sequence, $\ _t(o) \doteq \frac{(A_t j S_t; o)}{(A_t j S_t; O_t)}$ is the importance sampling ratio, and:

$$_t(o) \doteq R_{t+1} \quad R_t + u^{Q_t}(S_{t+1}; o) \quad Q_t(S_t; o); \tag{21}$$

where u^{Q_t} is defined in (13).

Theorem 4 *If $\hat{M}$ is communicating, Assumptions 1–3 and 5 hold, except for using $(t; s; o)$ instead of $(t; s; a)$, intra-option Differential Q-learning (Equations 19-21) converges, almost surely, Q_t to the set of solutions of (11) and (18), R_t to $\hat{r}$, and $r(\ _t; s)$ to $\hat{r}$ where $\ _t$ is a greedy policy w.r.t. Q_t.*

Appendix B. Proofs

B.1. Proof of Proposition 1

Note that the execution time of each option $o \ 2 \ O$ is the return of executing this option's policy in a stochastic shortest path MDP (SSP-MDP, Bertsekas 2007) with state space $S + f?g$ ($?$ is the "terminal" state of the SSP-MDP), action space A and transition function $\hat{p}$ satisfying:

$$\hat{p}(s^0; 1 \, j \, s; a) \doteq (1 \quad (s^0; o)) \sum_{r} p(s^0; r \, j \, s; a); \quad 8\, s; s^0 \,6\!=\, ?; a \,2\, A;$$

$$\hat{p}(?; 1 \, j \, s; a) \doteq \sum_{s^0} \quad (s^0; o) \sum_{r} p(s^0; r \, j \, s; a); \quad 8\, s \,6\!=\, ?; a \,2\, A;$$

$$\hat{p}(?; 0 \, j \, ?; a) \doteq 1; \quad 8\, a \,2\, A:$$

Also, note that by Assumption 5, the option's policy is a 'proper' policy (Bertsekas 2007) in the SSP-MDP. That is, 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 an SSP-MDP with state space $S + f?g$, action space A, and transition function $\hat{p}$ satisfying:

$$\hat{p}(s^0; r \, j \, s; a) \doteq (1 \quad (s^0; o)) p(s^0; r \, j \, s; a); \quad 8\, s; s^0 \,6\!=\, ?; a \,2\, A$$

$$\hat{p}(?; r \, j \, s; a) \doteq \sum_{s^0} \quad (s^0; o) p(s^0; r \, j \, s; a); \quad 8\, s \,6\!=\, ?; a \,2\, A$$

$$\hat{p}(?; 0 \, j \, ?; a) \doteq 1; \quad 8\, a \,2\, A:$$

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

B.2. Proof of Proposition 2

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

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

$$\begin{aligned}
& q(s, o) \\
&= \sum_{s', r, \tau} \hat{p}(s', r, \tau \mid s, o)(r - \tau r + \max_{o'} q(s', o')) \\
&= \sum_{s', r, \tau} \sum_a \pi(a \mid s, o) \sum_{s', r} p(s', r \mid s, a) \\
&\quad [\ \beta(s', o) \mathbb{I}(s = s', r = r, \tau = 1) + (1 - \beta(s', o)) \hat{p}(s', r - r, \tau - 1 \mid s', o)](r - \tau r + \max_{o'} q(s', o')) \\
&= \sum_a \pi(a \mid s, o) \sum_{s', r} p(s', r \mid s, a)\ \beta(s', o)(r - r + \max_{o'} q(s', o')) \\
&\quad + \sum_a \pi(a \mid s, o) \sum_{s', r} p(s', r \mid s, a)(1 - \beta(s', o)) \sum_{s', r, \tau} \hat{p}(s', r - r, \tau - 1 \mid s', o)](r - \tau r + \max_{o'} q(s', o')) \\
&= \sum_a \pi(a \mid s, o) \sum_{s', r} p(s', r \mid s, a)\ \beta(s', o)(r - r + \max_{o'} q(s', o')) \\
&\quad + \sum_a \pi(a \mid s, o) \sum_{s', r} p(s', r \mid s, a)(1 - \beta(s', o)) \sum_{s', r, \tau} \hat{p}(s', r, \tau \mid s', o)](r + r - (\tau + 1) r + \max_{o'} q(s', o')) \\
&= \sum_a \pi(a \mid s, o) \sum_{s', r} p(s', r \mid s, a)\ \beta(s', o)(r - r + \max_{o'} q(s', o')) \\
&\quad + \sum_a \pi(a \mid s, o) \sum_{s', r} p(s', r \mid s, a)(1 - \beta(s', o)) \left(r - r + \sum_{s', r, \tau} \hat{p}(s', r, \tau \mid s', o)(r - \tau r + \max_{o'} q(s', o')) \right) \\
&= \sum_a \pi(a \mid s, o) \sum_{s', r} p(s', r \mid s, a)\ \beta(s', o)(r - r + \max_{o'} q(s', o')) \\
&\quad + \sum_a \pi(a \mid s, o) \sum_{s', r} p(s', r \mid s, a)(1 - \beta(s', o))\ r - r + q(s', o) \\
&= \sum_a \pi(a \mid s, o) \sum_{s', r} p(s', r \mid s, a)(r - r + u^q(s', o))
\end{aligned}$$

Therefore any solution of (11) must be a solution of (12)-(13) and vice versa.

B.3. Proof sketch of Theorem 1-Theorem 4

In this section, we sketch the proof of Theorems 1-4. 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 B.8. 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 4. We defer this part to Appendix B.9. Given that the three assumptions are verified, 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 B.10.

B.4. Characterization of the Solution Set

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

We consider the set of solutions of q in the SMDP optimality equation ((11)) and

$$r^* = f(q); \tag{22}$$

where $f : S \times O \to \mathbb{R}$ satisfies Assumption 4. It is clear that (11) generalizes over (2) and (22) generalizes over (6), (9), and (18). And thus the characterization of Q_1 applies to the sets that action/option values in the aforementioned algorithms are claimed to converge to in Theorems 1–4.

It is known that if the SMDP is weakly communicating, (11) has r^* as its unique solution of r. For q, it has been shown by Schweitzer & Federgruen (1978) (we will refer to this work 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 of q in (11) is closed, unbounded, connected, and possibly non-convex. The next theorem characterizes Q_1.

Theorem 5 *If the SMDP is weakly communicating and Assumption 4 holds, Q_1 is non-empty, closed, bounded, connected, and possibly non-convex.*

Proof

First, Q_1 is non-empty. To see this point, note that for any solution of q in (11), q, $q + ce$ is also a solution for any $c \in \mathbb{R}$ and thus there must be a c such that (22) holds because $f(x + ce) = f(x) + cu$ for any $x; c$.

Q_1 is closed because the set of solutions of q in (11) is closed by S&F, the set of solutions of q in (22) is closed because f is Lipschitz and is thus continuous, and the intersection of two closed sets is closed.

Boundedness

We now show that Q_1 is bounded. For any $q \in Q_1$, let $v(s) \doteq \max_o q(s,o)$. Rewrite the option-value optimality equation (Equation (11)) using v instead of q, we have,

$$v(s) = \max_o \sum_{s',r,\tau} p(s',r,\tau \mid s,o)(r - \hat{r} + v(s')), \forall s \in S. \tag{23}$$

The above equation is known as the state-value optimality equation.

It is easy to verify that

$$q(s,o) = \sum_{s',r,\tau} p(s',r,\tau \mid s,o)(r - \hat{r} + v(s')). \tag{24}$$

Using this fact, rewrite Equation (22) using v instead of q, we have,

$$\hat{r} = f(r + Pv), \tag{25}$$

where

$$r(s,o) \doteq \sum_{s',r,\tau} p(s',r,\tau \mid s,o)(r - \hat{r}),$$
$$P(s,o,s') \doteq \sum_{r,\tau} p(s',r,\tau \mid s,o).$$

Denote the set of solutions of v in (23) by V. Denote the set of solutions of v in Equation (23) and Equation (25) by V_1. If V_1 is bounded, Q_1 is also bounded because any $q \in Q_1$ can be obtained from a solution of $v \in V_1$ with a linear operation in view of Equation (24).

In order to show boundedness, We will need the following two lemmas to proceed. These two lemmas are 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 B.5, B.6.

To formally state the Lemma 1, we will first introduce some definitions.

For any $\pi \in \Pi$, let P_π denote the $|S| \times |S|$ transition probability matrix under policy π. That is,

$$P_\pi(s,s') \doteq \sum_{o,r,\tau} \pi(o \mid s) p(s',r,\tau \mid s,o). \tag{26}$$

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

$$P_\pi^* \doteq \lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} P_\pi^i. \tag{27}$$

Because S is finite, the Cesaro limit exists and P_π^* is a stochastic matrix (has row sums equal to 1). Let $l_\pi(s) \doteq \sum_{o,s',r,\tau} \pi(o \mid s) p(s',r,\tau \mid s,o) l$. And let the fundamental matrix $Z_\pi \doteq (I - P_\pi^* + P_\pi^*)^{-1} = I + \lim_{n \to \infty} \sum_{k=1}^{n} (P_\pi^k - P_\pi^*)$.

Lemma 1 *If the SMDP is weakly communicating, v is a solution of (23) if and only if*

$$v(s) = \max_{\pi \in \Pi}[R_\pi - \rho_\pi \tau_\pi + P_\pi v](s), \forall s \in S. \tag{28}$$

In order to state Lemma 2 formally, we need the following definitions. Define the Bellman error for a state $s \in S$ given a policy $\pi \in \Pi$ and some $v \in \mathbb{R}^{|S|}$, $\delta_{v,\pi}(s)$, as follows:

$$\delta_{v,\pi}(s) \doteq [R_\pi - \rho_\pi \tau_\pi + P_\pi v - v](s).$$

Define R^* as the set of recurrent states for P_{π^*}. That is,

$$R^* \doteq \{s \mid P_{\pi^*}(s,s) > 0\}.$$

Let Π^* denote the set of optimal meta policies. Define R^* as the set of states that are recurrent under some optimal meta policy:

$$R^* \doteq \{s \mid s \in R_\pi \text{ for some policy } \pi \in \Pi^*\}.$$

By Theorem 3.2 (b) in S&F, there exists a policy $\pi^* \in \Pi^*$ such that $R_{\pi^*} = R^*$. For any $\pi \in \Pi^*$, let $n(\pi)$ be the number of recurrent classes for P_π. Further, define the least number of recurrent classes induced by any optimal meta policy that induces the set of recurrent states R^*:

$$n^* \doteq \min\{n(\pi) \mid \pi \in \Pi^*, R_\pi = R^*\}.$$

Denote Π^{n^*} as the set of policies that have R^* as their sets of recurrent states and that have n^* recurrent classes. That is,

$$\Pi^{n^*} \doteq \{\pi \in \Pi^* \mid R_\pi = R^*, n(\pi) = n^*\}.$$

Theorem 3.2 (d) by S&F shows that all policies within Π^{n^*} share the same collection of recurrent classes. Denote the collection of recurrent classes as $\{R_i \mid i = 1, 2, \ldots, n^*\}$. The following lemma shows that the solution set of (23) has n^* degrees of freedom.

Lemma 2 *If the SMDP is weakly communicating, suppose v and $v + x$ are both solutions of (23), then there exists n^* constants $y_1, y_2, \ldots, y_{n^*}$ such that*

$$x(s) = y_i, \forall i \in R_i, \quad i = 1, \ldots, n^* \tag{29}$$

$$x(s) = \max_{\pi \in \Pi^*}[R_\pi - \rho_\pi \tau_\pi](s) + \sum_{i=1}^{n^*} \Big(\sum_{s' \in R_i} P_{\pi^*}(s, s') \Big) y_i, \quad s \in S \setminus R^*, \tag{30}$$

$$y_i \geq [R_\pi - \rho_\pi \tau_\pi](s) + \sum_{i=1}^{n^*} \Big(\sum_{s' \in R_i} P_\pi(s, s') \Big) y_i, \quad i = 1, \ldots, n^*, s \in R_i, \pi \in \Pi^*. \tag{31}$$

For any $2 \{1;2; \ldots;n\}$, note that there exists a policy $(\ )$ such that R is the only one recurrent class under $(\ )$. To see this point, note that the SMDP is weakly communicating, and thus we can modify to obtain a new meta policy such that all states except for those in R are transient.

Based on the above observation and Lemma 2, for any $v \ 2 \ V$, we have for any given $2 \{1;\ldots;n\}$, there exists a $(\ )$, such that

$$y \quad \max_{s2R;} Z_{(\ )} b_{v;(\ )}(s) + \sum_{s^0 2R} P^{1}_{(\ )}(s;s^0)y$$
$$= \max_{s2R;} Z_{(\ )} b_{v;(\ )}(s) + y; \quad 8 \quad = 1; \quad ;n:$$

The first term $\max_{s2R;} Z_{(\ )} b_{v;(\ )}(s)$ is a constant given v and $(\ )$. Therefore we see that, for any other solution $v + x$ of (23), if y is arbitrarily large then $y; 8 \quad = 1; \quad ;n$ should also be arbitrarily large. This would violate the Lipschitz assumption on f. To see this point, let

$$\tilde{f}(v) \doteq f(r + Pv): \tag{32}$$

Let L be a Lipschitz constant of f. L is also a Lipschitz constant of $\tilde{f}$ because P is a stochastic matrix and is thus a non-expansion. Choose a $v_1 \ 2 \ V_1$ and a $v_2 \ 2 \ V_1$. a $v_2 \ 2 \ V_1$, denote $m = kv_1 \quad v_2k$. Choose a sufficiently large $c > 0$ such that $cu > L\,kv_1 + ce \quad v_2k = L\,kv_1 + ce \quad v_2 \quad cek = Lm$, where $v_2 \doteq v_2 + ce$. Given this choice of c, using $\tilde{f}(v_1) = \tilde{f}(v_2) = \hat{r}$ and $\tilde{f}(v_1 + ce) = \tilde{f}(v_1) + cu$, we have $\tilde{f}(v_1 + ce) \quad \tilde{f}(v_2) = \tilde{f}(v_1) + cu \quad \tilde{f}(v_2) = cu > L\,kv_1 + ce \quad v_2k$. This inequality suggests that $\tilde{f}$ is not Lipschitz continuous with a Lipschitz constant L and thus violates our assumption. Because the choice of is arbitrary, V_1 is upper bounded.

In addition, because the choice of is arbitrary, we have for any $2 \{1;\ldots;n\}$,

$$y \quad \max_{2\{1;\ldots;n\}} \max_{s2R;} Z_{(\ )} b_{v;(\ )}(s) + y$$

If y is chosen to be arbitrarily small then y should also be arbitrarily small for all $= 1; \quad ;n$ but again this is not allowed due to (25) for the same reason as mentioned in the previous paragraph. Therefore $y; 8 \quad 2 \{1;\ldots;n\}$ can not be arbitrarily small. Thus V_1 is lower bounded. Combining the upper bound and lower bound, V_1 is bounded. Therefore Q_1 is also bounded.

Connectedness

We now show that Q_1 is connected. To this end, again it is enough to show that V_1 is connected.

Define a function that takes a $v \ 2 \ V$ as input and produces an element in V_1 as output. Specifically, let $g : V \ ! \ V_1$ with $g(v) = v + xe$, where x is the solution of $\tilde{f}(v + xe) = \hat{r}$ and $\tilde{f}$ is defined in (32). Note that x is unique given v because $\hat{r} = \tilde{f}(v+xe) = \tilde{f}(v)+xu$ and thus $x = (\hat{r} \quad \tilde{f}(v))=u$.

We now show that g is Lipschitz continuous. Consider any $v_1; v_2 \in V$. 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

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

$$\begin{aligned} |x_1 \quad x_2| &= |\tilde{f}(v_1 + x_1 e) \quad \tilde{f}(v_2 + x_1 e)|/u \\ &\leq \|v_1 + x_1 e \quad v_2 \quad x_1 e\|/u \\ &= \|v_1 \quad v_2\|/u \end{aligned}$$

$$|g(v_1) \quad g(v_2)| = \|v_1 + x_1 e \quad v_2 \quad x_2 e\| \quad \|v_1 \quad v_2\| + |x_1 \quad x_2| \, \|e\| \leq (1 + L/u)\|v_1 \quad v_2\|.$$

Therefore g is Lipschitz continuous with Lipschitz constant $1 + L/u$.

Finally, because V is connected and the image of any continuous function on a connected set is connected, $g(V)$ is connected. Note that every point in $g(V)$ belongs to V_1 by definition. Every point in V_1 also belongs to $g(V)$. To see this point, pick any $v \in V_1$, we can see that $v \in V$ and that $g(v) = v$ (note that $x = 0$ given that $v \in V_1$). Thus $v \in g(V)$. Therefore $V_1 = g(V)$ is connected.

Given that V_1 is connected, Q_1 should also be connected because Q_1 is a linear transformation of V_1 (see Equation 24).

Non-convexity

We now show that both V_1 and Q_1 are not necessarily convex. We will show this point by constructing a counter-example, which involves the communicating MDP shown in the left sub-figure of Figure 2. The optimal reward rate for the MDP is 0.

Let $f(q) = \sum_{s,a} q(s,a)$. Such a choice of f satisfies the assumption on f in Theorem 5. Then by definition, for any $v \in V_1$, we have

$$f(r + Pv) = f.$$

For the three-state MDP considered here, $f = 0$.

$$\begin{aligned} f(r + Pv) &= \sum_{s,a} r(s,a) + p(s' \mid s,a)v(s') \\ &= 3 + 2v(1) + 3v(2) + v(3). \end{aligned}$$

Therefore,

$$2v(1) + 3v(2) + v(3) = 3.$$

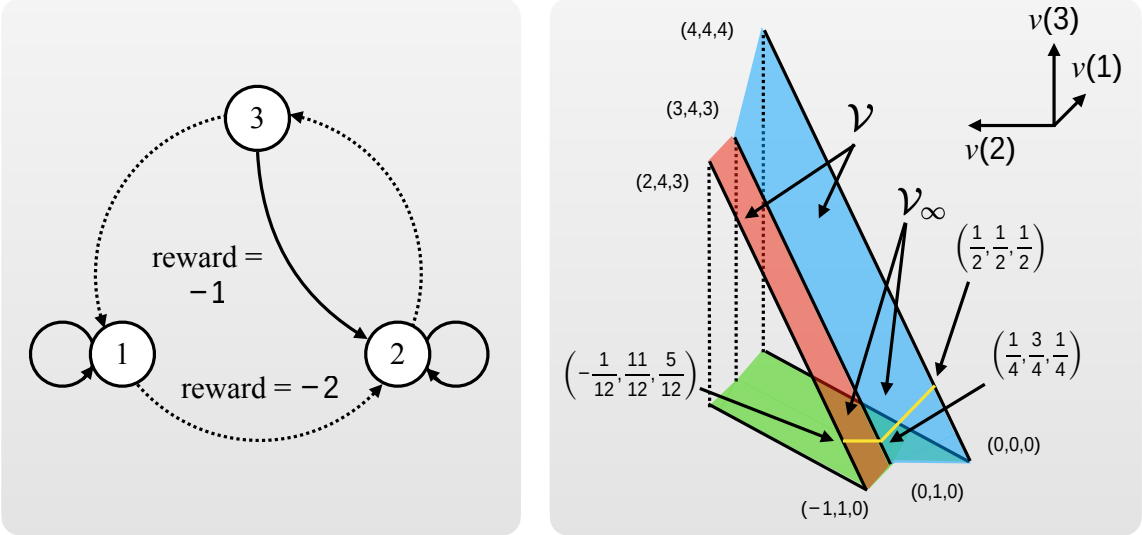

Figure 2: Illustration example. *Left*: the example MDP. There are three states marked by three circles respectively. There are two actions `solid` and `dashed`, both have deterministic effects. Taking action `solid` at state *3* results in a reward of 1. Taking action `dashed` at state *1* results in a reward of 2. All other rewards are 0. *Right*: a graphical explanation of $V; V_1$. The two yellow line segments together represent the solution set V_1, which is not convex. The red and blue regions together represent V.

In addition to the above equality, V_1 needs to satisfy the state-value optimality equation (Equation (23)). Therefore for any $v \in V_1$,

$$\begin{aligned} v(1) &= \max(v(1);\ 2 + v(2)); \\ v(2) &= \max(v(2); v(3)); \\ v(3) &= \max(v(1);\ 1 + v(2)); \end{aligned}$$

which implies

$$\begin{aligned} v(1) &\geq 2 + v(2); \\ v(2) &\geq v(3); \\ v(3) &= \max(v(1);\ 1 + v(2)): \end{aligned}$$

Therefore

$$\begin{aligned} V_1 = \{v \in \mathbb{R}^3 \mid & v(1) \geq 2 + v(2); v(2) \geq v(3); v(3) = \max(v(1);\ 1 + v(2)); \\ & 2v(1) + 3v(2) + v(3) = 3g: \end{aligned}$$

Graphically, V_1 corresponds to the two connected yellow line segments in the right sub-figure in Figure 2. From the figure, we see that V_1 is not convex.

Let s and d denote `solid` and `dashed` respectively. Consider any $q \in Q_1$, in view of (22),

$$q(1; s) + q(1; d) + q(2; s) + q(2; d) + q(3; s) + q(3; d) = 0:$$

In addition to the above equality, Q_1 needs to satisfy the action-value optimality equation (Equation (2)). Therefore for any $q \in Q_1$,

$$\begin{aligned}
q(1,s) &= 0 \quad 0 + \max(q(1,s), q(1,d)) \\
q(1,d) &= \quad 2 \quad 0 + \max(q(2,s), q(2,d)) \\
q(2,s) &= 0 \quad 0 + \max(q(2,s), q(2,d)) \\
q(2,d) &= 0 \quad 0 + \max(q(3,s), q(3,d)) \\
q(3,s) &= \quad 1 \quad 0 + \max(q(2,s), q(2,d)) \\
q(3,d) &= 0 \quad 0 + \max(q(1,s), q(1,d)),
\end{aligned}$$

which implies

$$\begin{aligned}
q(1,s) &\quad q(1,d) \\
q(1,d) &= \quad 2 + \max(q(2,s), q(2,d)) \\
q(2,s) &\quad q(2,d) \\
q(2,d) &= \max(q(3,s), q(3,d)) \\
q(3,s) &= \quad 1 + \max(q(2,s), q(2,d)) \\
q(3,d) &= \max(q(1,s), q(1,d))
\end{aligned}$$

Consider two solutions $q_1, q_2 \in Q_1$ defined as follows:

$$\begin{aligned}
q_1(1,s) &= \frac{1}{2}, q_1(1,d) = \quad \frac{3}{2}, \\
q_1(2,s) &= \frac{1}{2}, q_1(2,d) = \frac{1}{2}, \\
q_1(3,s) &= \quad \frac{1}{2}, q_1(3,d) = \frac{1}{2}, \\
q_2(1,s) &= \quad \frac{2}{3}, q_2(1,d) = \quad \frac{2}{3}, \\
q_2(2,s) &= \frac{4}{3}, q_2(2,d) = \frac{1}{3}, \\
q_2(3,s) &= \frac{1}{3}, q_2(3,d) = \quad \frac{2}{3}.
\end{aligned}$$

The midpoint of q_1 and q_2, $q \doteq 0.5q_1 + 0.5q_2$, satisfies

$$\begin{aligned}
q(1,s) &= \quad \frac{1}{12}, q(1,d) = \quad \frac{13}{12}, \\
q(2,s) &= \frac{11}{12}, q(2,d) = \frac{5}{12}, \\
q(3,s) &= \quad \frac{1}{12}, q(3,d) = \quad \frac{1}{12}.
\end{aligned}$$

Note that

$$\frac{5}{12} = q(2,d) \neq \max(q(3,s), q(3,d)) = \quad \frac{1}{12}.$$

Therefore q does not satisfy the action-value optimality equation ((2)) and $q \notin Q_1$. Thus Q_1 is not convex. ■

B.5. Proof of Lemma 1

Part 1: (23) $\Rightarrow$ (28)**:**

Choose any solution of (23), v, and choose any $\pi \in \Pi$. Using Theorem 3.1 (e) by S&F,

$$\begin{aligned} 0 &\le r - \rho^\pi \mathbf{1} + P v - v \\ (I - P)v &\le r - \rho^\pi \mathbf{1} \end{aligned}$$

(in weakly communicating SMDPs, the "L" set in Theorem 3.1 (e) is O).

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

$$\begin{aligned} (I - P^\pi)v &\le Z^\pi (r - \rho^\pi \mathbf{1}) \\ v &\le Z^\pi (r - \rho^\pi \mathbf{1}) + P^\pi v \end{aligned}$$

Because π can be any element in Π,

$$v(s) \le \sup_{\pi \in \Pi} [Z^\pi (r - \rho^\pi \mathbf{1}) + P^\pi v](s), \forall s \in S. \tag{33}$$

By Theorem 4.1 (c) in S&F, there exists a deterministic optimal policy π, which is apparently an element of Π such that

$$v = Z^\pi (r - \rho^\pi \mathbf{1}) + P^\pi v.$$

This result, along with (33) shows that

$$v(s) = \max_{\pi \in \Pi} [Z^\pi (r - \rho^\pi \mathbf{1}) + P^\pi v](s), \forall s \in S.$$

Part 1 is proven.

Part 2: (28) $\Rightarrow$ (23)**:**

Conversely, if v satisfies (28), define

$$\forall (s) \doteq \max_a \sum_{s', r, l} p(s', r, l \mid s, a)(r - l\rho^\pi + v(s')), \forall s \in S. \tag{34}$$

We first show that $\forall \quad v$. For any $\quad 2 \ ^{\wedge}$,

$$\begin{aligned}
\forall(s) &= \max_{o} \sum_{s^0;r;l} \not{p}(s^0; r; l \mid s; o)(r \quad l \not{\wedge} + v(s^0)) \\
&\quad [r \quad l \ \not{\wedge} + P \ v](s) && \text{because of the ``max''} \\
&\quad [r \quad l \ \not{\wedge} + P \ (Z \ (r \quad l \ \not{\wedge}) + P^{\ 1} v)](s) && \text{by (28)} \\
&= [(I + P \ Z \ )(r \quad l \ \not{\wedge}) + P^{\ 1} v](s) && \text{rearranging terms} \\
&= [(Z \ + P^{\ 1})(r \quad l \ \not{\wedge}) + P^{\ 1} v](s) && \text{by Equation 2.2 in S\&F} \\
&= [Z \ (r \quad l \ \not{\wedge}) + P^{\ 1} v](s): && \text{by Theorem 3.1 (a) in S\&F} \\
&= v(s) && \text{by (28)}
\end{aligned}$$

We now show that $\forall \quad v$, which, together with $\forall \quad v$, implies $\forall = v$.

Let $\quad$ be a deterministic policy achieving all jSj maxima in (34). Then we have

$$v \quad \forall = r \quad l \not{\wedge} + P \ v \tag{35}$$

$$(I \quad P \ )v \quad r \quad l \not{\wedge}: \tag{36}$$

Multiplying both sides by $P^{\ 1} \quad 0$, we have

$$P^{\ 1}(I \quad P \ )v \quad P^{\ 1}(r \quad l \not{\wedge}):$$

The l.h.s. is 0 because $P^{\ 1} P \ = P^{\ 1}$. The r.h.s. $\quad 0$ because

$$P^{\ 1}(r \quad l \not{\wedge}) \quad P^{\ 1}(r \quad l \ r(\ )) = 0;$$

where we use $\not{\wedge} \quad r(\ ); 8 \quad 2 \ ^{\wedge}$. Combining the above two inequalities, we have

$$P^{\ 1}(r \quad l \not{\wedge}) = 0: \tag{37}$$

Now we need to use the following lemma, which is essentially the same as Lemma 2.1 by S&F, except that the signs of inequalities are reversed. The proof follows the same arguments as those used in the proof of Lemma 2.1 by S&F.

Lemma 3 *Fix any policy* $\quad 2 \ ^{\wedge}$*, suppose that* $P^{\ 1} b = 0$ *and* $(I \quad P \ )x \quad b \quad 0$*, then* $(I \quad P^{\ 1})x \ Z \ b \quad 0$.

Let $x = v$, and $b = r \quad l \not{\wedge}$, we see that $P^{\ 1} b = 0$ because of (37) and $(I \quad P \ )x \quad b \quad 0$ because of (36). Using the above lemma, we have

$$v \quad Z \ (r \quad l \not{\wedge}) + P^{\ 1} v:$$

Inserting this inequality to (35), we have

$$\begin{aligned}
\Psi &= [r - \rho^* I + P v] \\
&\quad (r - \rho^* I + P [Z (r - \rho^* I) + P^{-1} v]) \\
&= [(I + P Z)(r - \rho^* I) + P^{-1} v] \\
&= [(Z + P^{-1})(r - \rho^* I) + P^{-1} v] && \text{by Equation 2.2 in S\&F} \\
&= [Z (r - \rho^* I) + P^{-1} v] && \text{by Theorem 3.1 (a) in S\&F} \\
&\quad \max_{\pi' \in \Pi^*} [Z (r - \rho^* I) + P^{-1} v] \\
&= v && \text{because } v \text{ satisfies (28):}
\end{aligned}$$

Combining $\Psi \geq v$ and $\Psi \leq v$, we have $\Psi = v$ and therefore $v(s) = \Psi(s) = \max_o \sum_{s',r,l} p(s', r, l \mid s, o)(r - \rho^* l + v(s'))$, $\forall s \in S$. And thus v is a solution of (23).

B.6. Proof of Lemma 2

Choose $\pi \in \Pi^*$. Because $v, v + x$ are both solutions of (23), using part (e)(2) of Theorem 3.1 by S&F, we have, $\forall s \in R$,

$$\begin{aligned}
v(s) &= [r - \rho^* I + P v](s) \\
[v + x](s) &= [r - \rho^* I + P (v + x)](s).
\end{aligned}$$

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

$$x(s) = [P^{-1} x](s) = \lim_{n \to \infty} [P^n x](s), \forall s \in R_k,$$

where d_k is the unique stationary distribution of the k-th recurrent class of P^π. This proves (29).

Because $v + x$ is a solution of (23), by Lemma 1, $\forall s \in S$,

$$\begin{aligned}
[v + x](s) &= \max_{\pi' \in \Pi^*} [Z (r - \rho^* I) + P^{-1} [v + x]](s) \\
x(s) &= \max_{\pi' \in \Pi^*} [Z (r - \rho^* I + P^{-1} v - v) + P^{-1} x](s) && \text{by Equation 2.2 in S\&F} \\
&= \max_{\pi' \in \Pi^*} [Z b_{v,\pi} + P^{-1} x](s) \\
&= \max_{\pi' \in \Pi^*} [Z b_{v,\pi}](s) + \sum_{k=1}^{n} \Big(\sum_{s' \in R_k} P^{-1}(s, s') \Big) y_k.
\end{aligned}$$

Using the above equality and (29), (30) and (31) hold.

B.7. Uniqueness of the Zero Solution

The other result we will need to use to show the 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 4 *If an SMDP is weakly communicating and all rewards are* 0*,* 0 *is the only element in* Q_1.

Proof Given a weakly communicating SMDP, by Lemma 1, any solution of the state-value optimality equation ((23)) satisfies

$$v(s) \quad \sum_{s^0 2 R} P^{\ 1}(s;s^0)v(s^0); 8\, s\ 2\ R\ ; \quad 2\ \hat{}\ ;$$

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

Pick an arbitrary $v\ \ 2\ V$ and an arbitrary policy $\ \ 2\ \hat{}$. For each recurrent class C under P, we have, by Lemma 1, $8\, s\ 2\ C$, $v\ (s)$ $\sum_{s^0 2 C} d^C(s^0)v\ (s^0)$, where d^C denotes the stationary distribution of $\ $ in the recurrent class C. The r.h.s. only involves class C because starting from a state $s\ 2\ C$ the MDP can not leave C. Because the choice of s is arbitrary, $\min_{s2C} v\ (s)$ $\sum_{s^0 2 C} d^C(s^0)v\ (s^0)$ $\min_{s2C} v\ (s)$: Thus $\sum_{s^0 2 C} d^C(s^0)v\ (s^0) = \min_{s2C} v\ (s)$. In addition, because $d^C(s) > 0$ for all $s\ 2\ C, v\ (s^0) = v(s); 8s; s^0\ 2\ C$.

Now for any $s; s^0\ 2\ R$, there must exist a $\ \ 2\ \hat{}\ = \hat{}$ such that there is a path from s to s^0 and a path from s^0 to s, because $s; s^0$ are in the same communicating class. Therefore $s; s^0$ are in the same recurrent class under P. Thus we conclude that $v\ (s) = v\ (s^0)$. Therefore $8\, s; s^0\ 2\ R$, $v\ (s) = v\ (s^0)$. The transient states values are uniquely determined by values of states in R. And in this case they are all equal to the values of states in the communicating class because all rewards are zero. Thus the solution set of v in the state-value optimality equation ((23)) is $fce : 8c\ 2\ Rg$.

Now consider the solution set of the option-value optimality equation ((11)). Let $v(s) = \max_o q(s; o)$, then (11) transforms to the state-value optimality equation. Therefore for any two solutions of q in (11), q_1 and q_2, $\max_o q_1(\ ; o) = \max_o q_2(\ ; o) + ce$ for some c. Furthermore, let q be any solution of q, $q\ (s; o_1) = q\ (s; o_2); 8\, s\ 2\ S; o_1; o_2\ 2\ O$ because $8\, s\ 2\ S; o\ 2\ O$:

$$\begin{aligned} q\ (s;o) &= \sum_{s^0;r} p(s^0; r\ j\ s; o) \max_{o^0} q\ (s^0; o^0) \\ &= \sum_{s^0;r} p(s^0; r\ j\ s; o) \max_{o^0} q\ (s; o^0) \\ &= \max_{o^0} q\ (s; o^0): \end{aligned}$$

Thus the solution set of q in the option-value optimality equation ((11)) is also $fce : 8c\ 2\ Rg$. Given (22), $cu = cf(e) = f(ce) = f\hat{}\ = 0$ and $u > 0$ implies that $c = 0$. Therefore 0 is the unique solution of q. The lemma is proved. ■

B.8. General RVI Q

We now start to present the General RVI Q algorithm.

Let $I \doteq f1; 2; \quad ; kg$ where k is a positive integer. Consider solving $r\ 2\ R$ and $q\ 2\ R^{jIj}$ in following equation

$$r(i) \quad r + g(q)(i) \quad q(i) = 0; 8\, i\ 2\ I \tag{38}$$

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 (38). This algorithm maintains an $|\mathcal{I}|$-dim vector of estimates $Q \in \mathbb{R}^{|\mathcal{I}|}$, and updates Q using

$$Q_{n+1}(i) \doteq Q_n(i) + \alpha_{\nu(n,i)} \big(R_n(i) - F_n(Q_n) + G_n(Q_n)(i) - Q_n(i) + \eta_n(i)\big) \text{ if } i \in Y_n, \tag{39}$$

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 $\mathbb{E}[R_n] = r, \forall\, n = 0,1,\ldots$,
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 satisfying $\mathbb{E}[F_n(Q)] = f(Q), \forall\, n = 0,1,\ldots$ where $f : \mathcal{I} \to \mathbb{R}$ is a function satisfying Assumption 4,
7. η_n is a sequence of random vectors of size $|\mathcal{I}|$.

We need the following assumptions in addition to Assumptions 1–4.

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 $c \in \mathbb{R}, x \in \mathbb{R}^{|\mathcal{I}|}$.*

Let

$$M_{n+1} \doteq R_n - r + G_n(Q_n) - g(Q_n) - (F_n(Q_n) - f(Q_n))e. \tag{40}$$

Let $\mathcal{F}_n \doteq \sigma(Q_0, \eta_0, M_1, \eta_1, M_2, \ldots, \eta_{n-1}, M_n)$ be the increasing family of σ-fields. By the above construction, $\{M_{n+1}\}$ is a martingale difference sequence w.r.t. $\mathcal{F}_n$. That is $\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = 0$ a.s., $\forall\, n \geq 0$.

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

We make the following assumption on η_n.

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

Assumption 9 *Equation 38 has a unique solution of r. That is, there exists a pair $r \in \mathbb{R}; q \in \mathbb{R}^{|\mathcal{I}|}$ satisfying (38) and if both $r_1; q_1$ and $r_2; q_2$ are solutions (38), $r_1 = r_2$.*

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

Define $\mathcal{Q}_\#$ to be the set of $q \in \mathbb{R}^{|\mathcal{I}|}$ satisfying (38) and

$$r_\# = f(q). \tag{41}$$

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 5 is assumed in Section 2 of Wan et al. (2021b). Assumptions 1–8 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 ((39)) converges, almost surely, $\mathcal{Q}_n$ to $\mathcal{Q}_\#$ and $f(\mathcal{Q}_n)$ to $r_\#$.*

Proof The proof would by a 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):

$$\underline{y}_t \doteq T_1(y_t) \quad y_t; \tag{42}$$

$$\underline{x}_t \doteq T_2(x_t) \quad x_t = T_1(x_t) \quad x_t + (r_\# \quad f(x_t))\, e; \tag{43}$$

where

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

Lemmas A.1 and A.3 by Wan et al. (2021) and their proofs still hold (by replacing r_1 with $r_\#$). Lemma A.2 by Wan et al. (2021) is replaced by the following one. The proof follows the same arguments as those for Lemma A.2.

Lemma 5 *The set of equilibrium points of (43) is $\mathcal{Q}_\#$.*

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

Lemma 6 *If $r(i) = 0; \forall i \in \mathcal{I}$, 0 is the globally asymptotically stable equilibrium for (43).*

Proof We have assumed that 0 is the only element in $\mathcal{Q}_\#$ in Assumption 11 and 0 is thus the unique equilibrium of (43). The rest follows the proof of Lemma A.4 by Wan et al. (2021) by replacing q_1 with 0. ■

Lemma 7 *$Q_\#$ is a compact connected internally chain transitive invariant set for the ODE (43). Furthermore, any set that contains points not in $Q_\#$ is not a compact connected internally chain transitive invariant set for the ODE (43).*

Proof According to Assumption 10, $Q_\#$ is closed and bounded and is thus compact. Assumption 10 also assumes that $Q_\#$ is connected. $Q_\#$ is an internally chain transitive invariant set for the ODE (43) because every element in $Q_\#$ is an equilibrium point of the ODE and $Q_\#$ is connected.

If a set contains a point $x \in \mathbb{R}^{|\mathcal{I}|}$ that is not in $Q_\#$, this set can not be internally chain transitive because x is "transient" and the trajectory of (43) can not be arbitrarily close to x at some arbitrarily large time step. ■

The last step is to show convergence of a synchronous version of (39)

$$Q_{n+1}(i) \doteq Q_n(i) + \alpha_{(n;i)} \big[R_n(i) - F_n(Q_n) + G_n(Q_n)(i) - Q_n(i) + \varepsilon_n(i) \big], \tag{44}$$

This result, together with Assumptions 2 and 3, guarantees convergence of the asynchronous update ((44)) by applying results from Section 7.4 by Borkar (2009)

Lemma 8 *Equation 44 converges a.s. Q_n to $Q_\#$ as $n \to \infty$.*

Proof The proof essentially follows that of Lemma A.5 with two changes. First, using Lemma 6 we can show that the ODE $\dot{x}_t = h_1(x_t) = g(x_t) - f(x_t)e - x_t$ 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 7 and Theorem 2 by Borkar (2009) together imply that Q_n must converge to $Q_\#$. ■

■

B.9. Verifying Assumption 9–Assumption 11

When casting General RVI Q to RVI Q-learning, (38) 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 4.

When casting General RVI Q to Differential Q-learning, (38) 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 R_0. It is clear that this new MDP, just like the original one, is communicating. Therefore Theorem 5 and Lemma 4 hold and thus Assumption 9–Assumption 11 hold.

When casting General RVI Q to inter-option Differential Q-learning, (38) becomes (11) 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 4 hold and thus Assumption 9–Assumption 11 hold.

When casting General RVI Q to intra-option Differential Q-learning, (38) becomes (12), which is the same as (11) by Proposition 2 for the reward shifted SMDP introduced in the previous paragraph.

B.10. Convergence of Reward Rates of Greedy Policies

Lemma 9 *Assume that the SMDP is weakly communicating, suppose Q_n converges to Q_1 almost surely, let π_n be a greedy policy w.r.t. Q_n, $r(\pi_n; s) \to r^*$ almost surely.*

Proof We will need the following lemma for the proof.

Lemma 10 *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 Q_1, consider $r(\pi_n; s)$ where π_n is a greedy policy w.r.t. Q_n. We show that $r(\pi_n; s)$ converges to r^* for all $s \in S$.

For any $\pi \in \Pi$, let P_π denote the $|S||O| \times |S||O|$ transition probability matrix under policy π. That is,

$$P_\pi(s,o,s',o') \doteq \sum_{r,l} p(s',r,l \mid s,o)\, \pi(o' \mid s'). \tag{45}$$

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

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

Because S is finite, the Cesaro limit exists and P_π^* is a stochastic matrix (has row sums equal to 1). Let $r(s,o) \doteq \sum_{s',r,l} p(s',r,l \mid s,o) r$ be the one-stage option reward and $l(s,o) \doteq \sum_{s',r,l} p(s',r,l \mid s,o) l$ be the one-stage option length. Let $r(\pi,s,o) \doteq \sum_{s',r,l} p(s',r,l \mid s,o) r(\pi,s')$ be the reward rate of policy π starting from s,o. By part (a) of Theorem 11.4.1 in Puterman (1994),

$$r(\pi_n,s,o) = \frac{P_n^* r(s,o)}{P_n^* l(s,o)}.$$

Thus we have,

$$\begin{aligned}
r(\pi_n,s,o) &= \frac{P_n^* r(s,o)}{P_n^* l(s,o)} \\
&= \frac{P_n^* (r + P_{\pi_n} Q_n - Q_n)(s,o)}{P_n^* l(s,o)} \\
&\quad \min_{s',o'} \frac{(r + P_{\pi_n} Q_n - Q_n)(s',o')}{l(s',o')} \\
&= \min_{s',o'} \frac{(TQ_n - Q_n)(s',o')}{l(s',o')},
\end{aligned}$$

where the inequality holds because of Lemma 10, and $TQ_n(s,o) \doteq r(s,o) + \sum_{s',r,l} p(s',r,l \mid s,o) \max_{o'} Q_n(s',o')$.

Because the SMDP is weakly communicating, consider a deterministic optimal policy $\pi^* \in \Pi$, $r(\pi^*,s) = r^*, \forall s \in S$. Therefore $r(\pi^*,s,o) = r^*$. Now we have, for any s,o,

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

The first inequality holds because of Lemma 10 and the second inequality holds because π_n is a greedy policy w.r.t. Q_n.

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

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

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 $Q_n - Q_1$ a.s., every point q in Q_1 satisfies $sp((Tq - q)/l) = 0$ because $(Tq - q)/l = \hat{r}e$ by (11). 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 $r(\pi_n; s; o) \to \hat{r}; \forall s; o$. By definition, $r(\pi_n; s) = \sum_o \pi_n(o|s) r(\pi_n; s; o)$. Therefore $r(\pi_n; s) \to \hat{r}$ ■

Appendix C. Empirical Results

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

We first consider the communicating MDP shown at the bottom of Figure 1. For this MDP, we apply Differential Q-learning with initial action values 0, initial reward rate estimate 3, and $\eta = 1$. The behavior policy 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 recorded the estimated action values and plotted the higher action value for each state in the figure. Figure 4(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 $f(q) = q(1; \texttt{dashed})$. A 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-value function converged to a point in the solution set (the black line segments), and 2) for different runs, the estimated action values generally converged to different points in the solution set.

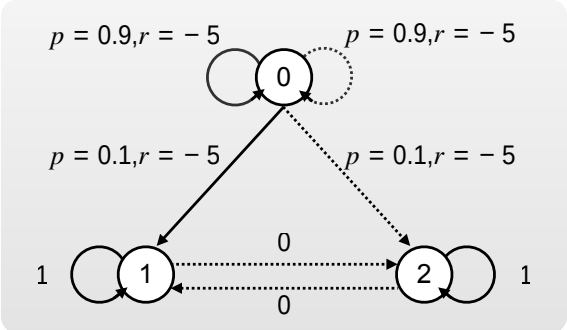

Figure 3: A weakly communicating MDP modified from the communicating MDP shown at the bottom of Figure 1 by adding a transient state 0.

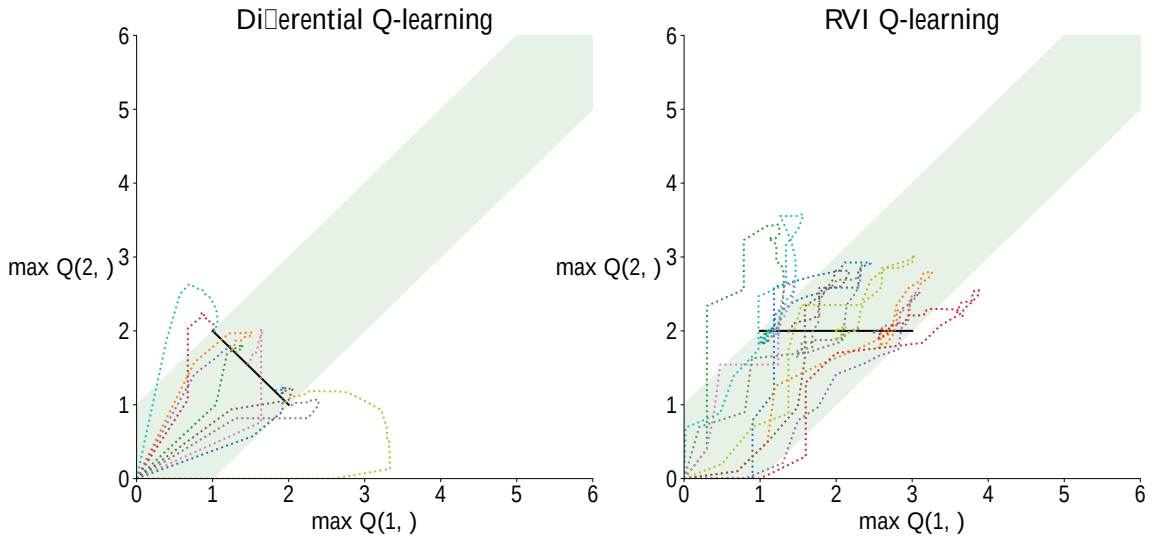

Figure 4: Evolution of estimated action values of Differential Q-learning and RVI Q-learning in the communicating MDP shown at the bottom of Figure 1. In each figure, the x and y axes are the higher estimated action value at state 1 and 2, respectively. The light green region marks the solution set of the action-value optimality equation (Equation (2)), and the black line segment marks the solution set Q_1. Each colored dotted trajectory marks the evolution of the estimated action values. Each trajectory starts from zero point and ends at some point on the black line segment.

We also applied both of the two algorithms, with the same parameter settings and initialization in a weakly communicating MDP (Figure 3), which is just the communicating MDP plus a transient state. In the transient state, taking both `solid` and `dashed` actions stays at the transient state with probability 0.9. The MDP moves to state 1 with probability 0.1 given action `solid` and to state 2 with probability 0.1 given action `dashed`. The reward starting from state 0 is always -5. The starting state is 0. Because the agent could spend different amounts of time in the transient state for different runs, the agent may enter the communicating set, which contains states 1 and 2, with different action values associated with state 0.

The solution set of Differential Q-learning depends on the action values associated with the transient states when entering the communicating class. Therefore in the figure, the points that the estimated action-value function converged to, corresponding to different runs, are not in a line. Nevertheless, the estimated action-value function in all runs converged to the green region, which corresponds to the solution set of the action-value optimality equation.

On the other hand, The solution set of RVI Q-learning with the choice of the reference function $f(q) = q(1, \texttt{dashed})$ does not depend on the action values associated with states in the communicating class when entering the class. Therefore the solution set did not vary across different runs. Note that if we chose $f(q) = q(0, \texttt{dashed})$, then again the solution set of RVI Q-learning has that dependence.

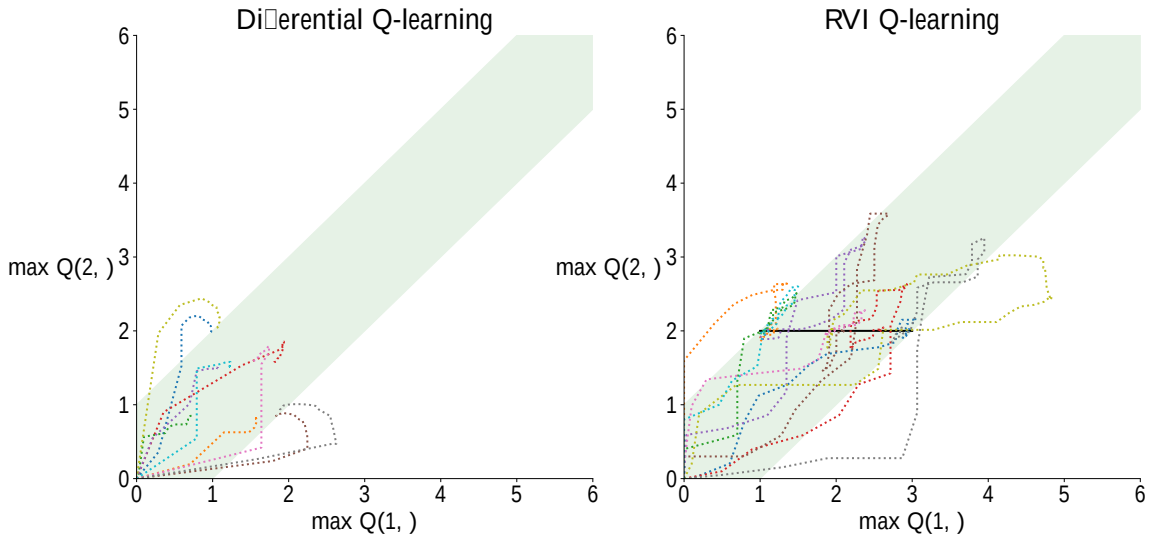

Figure 5: Evolution of estimated action values of Differential Q-learning and RVI Q-learning in the weakly communicating MDP shown in Figure 3.

Appendix D. Gosavi's (2004) Convergence Result Is Incorrect

The convergence result of Gosavi's proposed algorithm is presented in Theorem 2 of his paper. In the proof of the theorem, they used Borkar's two-time scale stochastic approximation result to prove the convergence of the proposed algorithm. Specifically, they argued that their algorithm is a special case of the general class of algorithms considered in Borkar's result. As Gosavi quotes, "Note that the Eqs. (48) and (49) for SMDPs form a special case of the general class of algorithms (29) and (30) analyzed using the lemma given in Section 5.1.1. " However, a closer look at these equations shows that equation (49) is not a special case of equation (30). Note that because ^{k} is a scalar, y^{k} only has one element and thus the f function in equation (30) does not vary across different state-option pairs. However, this is not true for the f function in equation (49). It appears to us that there is no simple fix for this issue.