An Analysis of Action-Value Temporal-Difference Methods That Learn State Values

Brett Daley, Prabhat Nagarajan, Martha White, Marlos C. Machado

Keywords: TD learning, QV-learning, Dueling DQN, advantage estimation.

Summary

The hallmark feature of temporal-difference (TD) learning is bootstrapping: using value predictions to generate new value predictions. The vast majority of TD methods for control learn a policy by bootstrapping from a single action-value function (e.g., Q-learning and Sarsa). Significantly less attention has been given to methods that bootstrap from two asymmetric value functions: i.e., methods that learn state values as an intermediate step in learning action values. Existing algorithms in this vein can be categorized as either QV-learning or AV-learning. Though these algorithms have been investigated to some degree in prior work, it remains unclear if and when it is advantageous to learn two value functions instead of just one—and whether such approaches are theoretically sound in general. In this paper, we analyze these algorithmic families in terms of convergence and sample efficiency. We find that while both families are more efficient than Expected Sarsa in the prediction setting, only AV-learning methods offer any major benefit over Q-learning in the control setting. Finally, we introduce a new AV-learning algorithm called Regularized Dueling Q-learning (RDQ), which significantly outperforms Dueling DQN in the MinAtar benchmark.

Contribution(s)

1. We prove the expected contraction of QV-learning for on-policy prediction.
   Context: Wiering (2005) introduced the QV-learning algorithm, but omitted a convergence proof. To our knowledge, there is no published convergence proof of QV-learning to date.
2. We raise the issue that QVMAX, the main off-policy control variant of QV-learning, is biased. We empirically demonstrate that such bias can significantly impact performance. We then introduce a new, unbiased algorithm, BC-QVMAX, and empirically demonstrate that it converges to a similar solution to that of Q-learning in our considered settings.
   Context: Wiering & van Hasselt (2009) introduced QVMAX without theoretical justification, heuristically mirroring the Q-learning update. Its bias has not been identified until now. Our empirical results were obtained with parametric MDP experiments that we designed; they should be interpreted only as anecdotal evidence for the convergence of BC-QVMAX.
3. In the context of AV-learning algorithms, we formalize a tabular version of Dueling DQN, and we introduce a new algorithm, Regularized Dueling Q-learning (RDQ). RDQ addresses the identifiability issue of the naive dueling decomposition by using an l_2 penalty instead of subtracting the mean advantage. We empirically demonstrate that, given the same network architecture, RDQ significantly outperforms Dueling DQN in the MinAtar domain.
   Context: Wang et al. (2016) introduced Dueling DQN from the perspective of an improvement to the neural network architecture used by DQN. RDQ is the result of relaxing the semantics behind estimating $Q(s, a)$ through two value functions. Instead of generating $Q(s, a)$ from approximations of $V(s)$ and $A(s, a)$, RDQ searches for the closest point on the hyperplane defined by two arbitrary functions that sum to $Q(s, a)$. We used tuned versions of DQN and Dueling DQN as baselines for MinAtar (Obando-Ceron & Castro, 2021).

An Analysis of Action-Value Temporal-Difference Methods That Learn State Values

Brett Daley[1,2,†], **Prabhat Nagarajan**[1,2,†], **Martha White**[1,2,3], **Marlos C. Machado**[1,2,3]
{brett.daley,nagarajan,whitem,machado}@ualberta.ca

[1]**Department of Computing Science, University of Alberta, Canada**
[2]**Alberta Machine Intelligence Institute (Amii)**
[3]**Canada CIFAR AI Chair**

[†] Co-first authors.

Abstract

The hallmark feature of temporal-difference (TD) learning is bootstrapping: using value predictions to generate new value predictions. The vast majority of TD methods for control learn a policy by bootstrapping from a single action-value function (e.g., Q-learning and Sarsa). Significantly less attention has been given to methods that bootstrap from two asymmetric value functions: i.e., methods that learn state values as an intermediate step in learning action values. Existing algorithms in this vein can be categorized as either QV-learning or AV-learning. Though these algorithms have been investigated to some degree in prior work, it remains unclear if and when it is advantageous to learn two value functions instead of just one—and whether such approaches are theoretically sound in general. In this paper, we analyze these algorithmic families in terms of convergence and sample efficiency. We find that while both families are more efficient than Expected Sarsa in the prediction setting, only AV-learning methods offer any major benefit over Q-learning in the control setting. Finally, we introduce a new AV-learning algorithm called Regularized Dueling Q-learning (RDQ), which significantly outperforms Dueling DQN in the MinAtar benchmark.

1 Introduction

Reinforcement learning (RL) is the study of decision-making agents that interact with their environment to maximize a notion of cumulative reward. Temporal-difference (TD) learning (Sutton, 1988) is among the most widely used classes of RL algorithms. Like dynamic programming (Bellman, 1957), TD methods learn one or more value functions to guide policy improvement. However, unlike dynamic programming, TD approaches do not assume access to or knowledge of the environment's dynamics. As such, value-function estimates must be generated iteratively through direct interaction with the MDP.

A key feature of TD is bootstrapping: constructing new value predictions from other value predictions. How such predictions should be integrated for the purposes of bootstrapping to learn as efficiently as possible has been the subject of much RL research. Classic methods like Q-learning (Watkins, 1989), Sarsa (Rummery & Niranjan, 1994), and Expected Sarsa (John, 1994) all learn action-value functions, and differ only by their bootstrapped targets, but exhibit dramatically different properties in terms of risk aversion, bias-variance trade-off, and convergence.

The vast majority of TD algorithms that attempt to learn good control policies—including the three algorithms mentioned above—rely on just a single action-value function, denoted by $Q(s, a)$ where

(s, a) is a state-action pair. A far less common paradigm is to jointly learn a secondary state-value function, $V(s)$, in the process of learning $Q(s, a)$. That is to say, such TD methods learn state values as an intermediate step in learning action values.[1]

We broadly categorize such methods as either *QV-learning* or *AV-learning*.[2] In QV-learning, the agent directly estimates $Q(s, a)$ and $V(s)$, with some bootstrapping between the two value functions. For instance, the classic QV-learning method (Wiering, 2005) essentially replaces the bootstrap term of Expected Sarsa with a learned approximation generated by TD(0) (Sutton, 1988). Conversely, AV-learning methods make use of the advantage decomposition (Baird, 1993), which breaks down the action value as $Q(s, a) = V(s) + \mathrm{Adv}(s, a)$. In contrast to QV-learning, the two value functions do not directly bootstrap from one another, but rather bootstrap from the composite action-value function, $Q(s, a)$. AV-learning techniques were popularized in deep RL by the dueling network architecture (Wang et al., 2016). Analyzing the behavior of its underlying algorithm, Dueling Q-learning, is important to understand how these networks learn.

In spite of past empirical evidence supporting QV-learning (e.g., Sabatelli et al., 2020; Modayil & Abbas, 2023) and AV-learning (Baird, 1993; Wang et al., 2016; Tang et al., 2023), the exact circumstances and mechanisms behind these improvements remain unclear. After all, estimating two value functions instead of one inherently requires more parameters to be learned. On the other hand, it seems that there are opportunities for synergistic information sharing between the value functions which could potentially explain the observed sample-efficiency gains. Still, it is unknown when, or how reliably, such gains can be obtained. Furthermore, several of these methods—particularly in the QV-learning family—are missing formal convergence guarantees.

Toward an understanding of when and how state-value estimation can be leveraged to learn action values more efficiently (and soundly), we investigate the theoretical underpinnings of both QV-learning and AV-learning algorithms and conduct experiments to isolate some of their unique properties. We show that QV-learning can be more efficient than Expected Sarsa for on-policy prediction and that the performance gap increases with the cardinality of the action set. However, QV-learning's main extension for off-policy control, QVMAX (Wiering & van Hasselt, 2009), suffers from bias; even when we correct this bias, we show empirically that its sample efficiency is surpassed by that of Q-learning. In contrast, we find that AV-learning methods work much more consistently for control problems overall and easily outperform Q-learning.

We also discuss the implicit assumptions underlying Dueling Q-learning and introduce an algorithm called Regularized Dueling Q-learning (RDQ) that converges to a different pair of value functions. An advantage of RDQ is that it easily extends to function approximation, and we show that it is significantly more sample efficient than Dueling DQN (Wang et al., 2016) in the five MinAtar games (Young & Tian, 2019) when using the same network architecture and hyperparameters. All of our experiment code is available online.[3] Our results help to characterize the nuanced distinction between these two algorithm families, and ultimately motivate state-value learning as a sound and effective way to accelerate action-value learning.

2 Background

We formalize the RL problem as a finite Markov decision process (MDP) described by $(\mathcal{S}, \mathcal{A}, p, \mathcal{R}, \gamma)$. At each time step $t \geq 0$, the agent observes the environment state, $S_t \in \mathcal{S}$, and executes an action, $A_t \in \mathcal{A}$. The environment consequently transitions to a new state, $S_{t+1} \in \mathcal{S}$, and returns a reward, $R_{t+1} \in \mathcal{R}$, with probability $p(S_{t+1}, R_{t+1} \mid S_t, A_t)$.

[1] Note that this precise definition of learning state and action values excludes double action-value methods such as Double Q-learning (van Hasselt, 2010), which are beyond the scope of this paper.

[2] We introduce the term *AV-learning* to refer to value-based RL methods that learn both advantage and state-value functions. AV-learning should not be confused with VA-learning (Tang et al., 2023), a specific algorithmic instance of AV-learning. Although VA-learning is inspired by Dueling DQN, it is not the exact tabular analog of it, which we derive in Section 4.1.

[3] https://github.com/brett-daley/reg-duel-q

In the policy-evaluation setting, the agent's goal is to learn the action-value function: the expected discounted return achieved by each state-action pair (s, a) under a fixed agent behavior. The agent's behavior is defined by a policy, a mapping from states to distributions over actions. Letting $G_t = \sum_{i=0}^{\infty} \gamma^i R_{t+1+i}$ be the discounted return at time t, the action-value function for policy π is defined as

$$q_\pi(s, a) = \mathbb{E}_\pi[G_t \mid (S_t, A_t) = (s, a)],$$

where the expectation $\mathbb{E}_\pi$ indicates that actions are sampled according to π. For any policy π, the corresponding action-value function q_π uniquely solves the Bellman equation (Bellman, 1957) for action values:

$$q_\pi(s, a) = \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r|s, a) \left(r + \gamma \sum_{a' \in \mathcal{A}} \pi(a'|s')\, q_\pi(s', a') \right). \tag{1}$$

In the control setting, the agent's goal is to find an *optimal* policy, π_*: one such that $q_{\pi_*}(s, a) \geq q_\pi(s, a)$ for every policy π and every state-action pair (s, a). Every optimal policy has the same action-value function, q_*, which uniquely solves the action-value Bellman optimality equation:

$$q_*(s, a) = \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r|s, a) \left(r + \gamma \max_{a' \in \mathcal{A}} q_*(s', a') \right).$$

Temporal-difference (TD) methods for control estimate action-value functions, either q_π or q_*, from sample-based interaction with the environment. Given a transition $(S_t, A_t, R_{t+1}, S_{t+1})$, the agent then conducts an incremental update to its estimate of the action-value function, $Q(S_t, A_t)$. In the *off-policy* setting, the agent is assumed to select actions with probability $b(A_t|S_t)$, where b is a behavior policy which differs from the target policy, π.

Most of the algorithms considered in this paper are off-policy methods. They learn an action-value function $Q(s, a)$ that estimates either $q_\pi(s, a)$ or $q_*(s, a)$ using samples obtained from executing policy b in the environment. For example, Expected Sarsa (John, 1994) is an off-policy TD method that uses explicit knowledge of the target policy π to approximate the solution to Eq. (1) from sampled data:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left(R_{t+1} + \gamma \sum_{a' \in \mathcal{A}} \pi(a'|S_{t+1}) Q(S_{t+1}, a') - Q(S_t, A_t) \right), \tag{2}$$

where $\alpha \in (0, 1]$ is the step size of the update. We revisit Expected Sarsa as an important baseline several times in this paper because it generalizes a number of fundamental bootstrapping algorithms (van Hasselt, 2011). For instance, when the target policy is greedy with respect to Q, the expectation in Eq. (2) becomes equivalent to $\max_{a' \in \mathcal{A}} Q(s', a')$, yielding the Q-learning algorithm:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left(R_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q(S_{t+1}, a') - Q(S_t, A_t) \right), \tag{3}$$

which will be useful for the control case in Section 3.2.

2.1 QV-learning

We now begin to discuss methods that learn state values as a means to estimate action values. The state-value function, v_π, is defined analogously to q_π as

$$v_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s]$$

and uniquely solves the Bellman equation

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r|s, a) \left(r + \gamma\, v_\pi(s') \right). \tag{4}$$

The state-value function is related to the action-value function by $v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s,a)$. This implies that an approximation of $v_\pi(s')$ can be substituted in Eq. (1) to avoid the summation over actions. QV-learning (Wiering, 2005) is an on-policy TD algorithm based on this concept. The idea is to use TD(0) to independently learn $V(s) \approx v_b(s)$, while concurrently bootstrapping from these estimated state values to learn $Q(s,a) \approx q_b(s,a)$. The specific value-function updates for QV-learning become

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\Big(R_{t+1} + \gamma V(S_{t+1}) - Q(S_t, A_t)\Big)\,, \tag{5}$$

$$V(S_t) \leftarrow V(S_t) + \alpha\Big(R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\Big)\,. \tag{6}$$

Wiering (2005, Sec. 2) is explicit that Q is always updated before V in each iteration. Additionally, it is commonly assumed that both updates share the same step size, α, as is shown above. We analyze this algorithm in Section 3.1.

The main variant of QV-learning for off-policy control is QVMAX (Wiering & van Hasselt, 2009), which essentially substitutes the max operator from Q-learning into Eq. (6) in an attempt to induce convergence to q_* instead of q_b:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\Big(R_{t+1} + \gamma V(S_{t+1}) - Q(S_t, A_t)\Big)\,, \tag{5}$$

$$V(S_t) \leftarrow V(S_t) + \alpha\Big(R_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q(S_{t+1}, a') - V(S_t)\Big)\,. \tag{7}$$

The update to $Q(S_t, A_t)$ remains the same as Eq. (5) and is still conducted prior to Eq. (7) at each time step. We analyze this algorithm in Section 3.2.

2.2 Dueling Q-learning

The basic idea behind Dueling Q-learning (Wang et al., 2016) is to decompose the action-value function, $Q(s,a)$, into a state-value function, $V(s)$, and an advantage function, $\mathrm{Adv}(s,a)$. The particular decomposition used by Wang et al. (2016) is

$$Q(s,a) = V(s) + \mathrm{Adv}(s,a) - \frac{1}{|\mathcal{A}|} \sum_{a' \in \mathcal{A}} \mathrm{Adv}(s,a')\,, \tag{8}$$

where the last term is subtracted to make the solution unique (we further discuss this point in Section 4.2). The updates to $V(s)$ and $\mathrm{Adv}(s,a)$ are then derived implicitly from Q-learning by reframing its update as stochastic semi-gradient descent. Let $\boldsymbol{\theta}$ be the vector of all learnable value-function parameters involved in Eq. (8). We therefore write $Q(s,a;\boldsymbol{\theta})$ for the action-value estimate of state-action pair (s,a). Additionally, let $\boldsymbol{\theta}^-$ be the target parameters for bootstrapping (Mnih et al., 2015). The Q-learning update in Eq. (3) is rewritten as the following squared-error loss minimization:

$$\begin{aligned}\boldsymbol{\theta} &\leftarrow \boldsymbol{\theta} - \alpha\, \frac{1}{2}\, \nabla_{\boldsymbol{\theta}} \left(R_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q(S_{t+1}, a'; \boldsymbol{\theta}^-) - Q(S_t, A_t; \boldsymbol{\theta})\right)^2 \\ &= \boldsymbol{\theta} + \alpha \left(R_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q(S_{t+1}, a'; \boldsymbol{\theta}^-) - Q(S_t, A_t; \boldsymbol{\theta})\right) \nabla_{\boldsymbol{\theta}}\, Q(S_t, A_t; \boldsymbol{\theta})\,. \end{aligned} \tag{9}$$

Applying the chain rule to $\nabla_{\boldsymbol{\theta}}\, Q(S_t, A_t; \boldsymbol{\theta})$ in Eq. (9) according to the value decomposition in Eq. (8) yields the final dueling update. In practice, the decomposition in Eq. (8) is represented by a branching and merging neural network, and the chain rule is handled by automatic differentiation. In Section 4.1, we instead apply the chain rule to Eq. (8) manually, assuming tabular value functions, to derive a Dueling Q-learning update without function approximation. This allows us to gain further insight into this fundamental algorithm beyond the deep RL setting.

3 QV-learning Algorithms

In this section, we investigate QV-learning algorithms, which jointly learn an action-value function $Q(s, a)$ and a state-value function $V(s)$. Although the fundamental QV-learning algorithm has existed for around 20 years and several studies have pointed to its effectiveness (e.g., Wiering, 2005; Wiering & van Hasselt, 2007; Sabatelli et al., 2020; Modayil & Abbas, 2023), empirical and theoretical analysis has still been limited.

3.1 On-Policy Prediction

In this subsection, we conduct a focused experiment to determine when QV-learning improves performance versus a single action-value function. We then conclude with an expected convergence analysis of QV-learning.

Throughout this paper, we will revisit Expected Sarsa as a prototypical baseline for TD learning of action values. Recall that Expected Sarsa's update rule is

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\Big(R_{t+1} + \gamma \underbrace{\sum_{a' \in \mathcal{A}} \pi(a'|S_{t+1}) Q(S_{t+1}, a')}_{\approx v_\pi(S_{t+1})} - Q(S_t, A_t)\Big). \tag{2}$$

The comparison to Expected Sarsa is natural because QV-learning can be seen as learning an approximation, $V(S_{t+1})$, of the expected next action value, which we have underscored above. This provides an indication as to when QV-learning may have an advantage over Expected Sarsa; as the action-space cardinality $|\mathcal{A}|$ of the MDP grows, Expected Sarsa requires many more sample interactions to estimate all of the action values in the expectation, whereas QV-learning can simply learn $V(S_{t+1})$ with a small number of samples that is roughly insensitive to the number of actions, $|\mathcal{A}|$.

To test this, we design a parametric MDP experiment where the environment has 4 states and $|\mathcal{A}|$ actions. Taking any action in any state triggers a transition to a random state (possibly the same one) with equal probability. The agents' behavior policy is uniform random in every state. Agents receive a reward of $+1$ whenever the first action, a_0, is taken, and a reward of 0 otherwise.

Because the MDP dynamics are state-invariant, $v_b(s)$ is constant for all $s \in \mathcal{S}$. Solving the state-value Bellman equation, Eq. (4), yields $v_b(s) = 1 \,/\, \big((1-\gamma)\,|\mathcal{A}|\big)$. It follows that $q_b(s, a_i) = \mathbf{1}_{\{i=0\}} + \gamma \,/\, \big((1-\gamma)\,|\mathcal{A}|\big)$ for all $a_i \in \mathcal{A}$, where $\mathbf{1}$ is the indicator function. We set $\gamma = 0.99$.

We evaluate the agents' performance by measuring the root mean square (RMS) prediction error, $\|q_b - Q\|_2$, as a function of the number of environment interactions. We normalize the errors by expressing them as a percentage of the initial RMS error, $\|q_b - Q_0\|_2$, where Q_0 is initialized with zeros. We compare Expected Sarsa and QV-learning, training both agents for 20,000 time steps. In Figure 1 (left), we plot the learning curves for the case where $|\mathcal{A}| = 18$, the largest instantiation of our action set that we consider here. We then average the results over 100 independent trials, with shading indicating 95% confidence intervals. The agents' step sizes are chosen from a natural-logarithmic grid search of 61 values over the interval $(0, 1]$, similarly to the random walk experiment by Sutton & Barto (2018, Sec. 12.1). The selection criterion is to minimize the area under the curve (AUC), which we normalized as a percentage. Specifically, an AUC of 100% indicates that the initial error did not change at all, whereas a lower percentage indicates faster average progress towards the fixed point. In Figure 1 (center), we plot the corresponding AUCs for each step size tested by the sweep. The dashed horizontal line corresponds to the smallest AUC achieved by each agent.

To investigate the scalability of these methods, we repeat this experiment setup for $|\mathcal{A}| \in \{2, 6, 10, 14, 18\}$. We then plot the best AUC achieved in each problem instance in Figure 1 (right). The slope of the resulting lines roughly correspond to the scalability of the algorithms. We note that QV-learning's line is more horizontal, indicating better scaling to large action-space cardinalities, as we hypothesized earlier. However, the trade-off for this improved sample efficiency is a much noisier update, which can be seen in Figure 1 (left); if we were to train for longer, Ex-

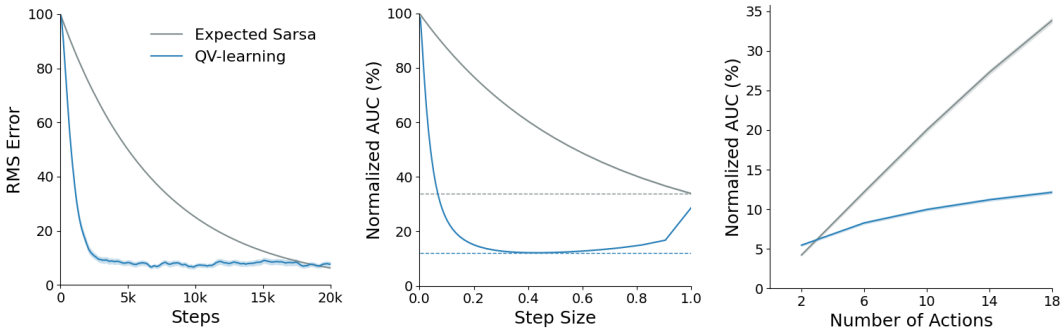

Figure 1: On-policy prediction performance of QV-learning compared to Expected Sarsa in the 4-state MDP. (Left) Root mean square (RMS) prediction error versus the number of learning steps, after optimizing the step size, $\alpha \in (0, 1]$. (Center) Area under the learning curve (AUC), normalized as a percentage, for each step size. The dashed lines correspond to the AUCs for the learning curves in the left subplot. (Right) The smallest AUC obtained across all step sizes, as the number of MDP actions is increased. Results are averaged over 100 independent trials. Shading represents 95% confidence intervals.

pected Sarsa would eventually obtain a more accurate solution. This indicates that QV-learning is perhaps best suited for cases where learning a satisfactory—but possibly imperfect—value function is desired within a small number of samples.

These results do not, however, suggest that QV-learning fails to converge to the correct pair of on-policy value functions, q_b and v_b. The noisy behavior observed in Figure 1 (left) is primarily due to the fact that the step size is not annealed. The following theorem states that, in expectation, the joint update of QV-learning is a contraction mapping when we represent the two value functions as a concatenated vector of size $|\mathcal{S}| + |\mathcal{S} \times \mathcal{A}|$.

Theorem 3.1 (QV-learning Contraction)**.** *The expected QV-learning update corresponds to an affine joint operator* $H\colon \left[\begin{smallmatrix} \boldsymbol{q} \\ \boldsymbol{v} \end{smallmatrix}\right] \mapsto \boldsymbol{b} + \boldsymbol{A}\left[\begin{smallmatrix} \boldsymbol{q} \\ \boldsymbol{v} \end{smallmatrix}\right]$*, where* $\boldsymbol{b} = \left[\begin{smallmatrix} \boldsymbol{r} \\ \boldsymbol{E}_b \boldsymbol{r} \end{smallmatrix}\right]$ *and* $\boldsymbol{A} = \gamma \left[\begin{smallmatrix} \boldsymbol{0} & \boldsymbol{P} \\ \boldsymbol{0} & \boldsymbol{E}_b \boldsymbol{P} \end{smallmatrix}\right]$*. The operator* H *is a contraction mapping with its unique fixed point equal to* $\left[\begin{smallmatrix} \boldsymbol{q}_b \\ \boldsymbol{v}_b \end{smallmatrix}\right]$.

Proof. See Appendix A.1. □

The significance of this result is that it shows that both of QV-learning's value functions make progress (on average) towards their respective fixed points, without requiring V to converge first. This is in contrast to less formal convergence arguments for QV-learning, which may rely on the assumption that TD(0) would first converge to v_b based on well-established convergence results, and then the update rule in Eq. (5) would be able to bootstrap from it to extract q_b. The issue with this two-timescale approach is that V may never exactly equal v_b after any finite amount of time, meaning that Q would still incur some bootstrapping bias—a caveat that is directly addressed by considering the joint operator space, as we have done here.

The fact that H is a contraction mapping implies that both Q and V converge to their respective fixed points even when updates are asynchronous, under the additional (but standard) technical assumptions that the step size α is appropriately annealed and the conditional variances of the updates are bounded (see Bertsekas & Tsitsiklis, 1996, Prop. 4.4). We note that the latter assumption is automatically satisfied by our MDP definition, which has finite sets of states, actions, and rewards. Finally, we remark that our analysis considers only the prediction setting in which the behavior policy b is fixed; for the control case, we would likely need to invoke the "greedy in the limit with infinite exploration" (GLIE) assumption (Singh et al., 2000) to prove eventual convergence to an optimal policy, but we leave this as an open problem.

3.2 Off-Policy Control

The previous experiment shows that bootstrapping from state values when learning action values can be much more efficient than directly learning the action values. Unfortunately, we cannot leverage this same technique for off-policy control, as this would require a model-free method for estimating v_* directly from environment samples.

To circumvent this, Wiering & van Hasselt (2009) introduced an off-policy variant of QV-learning called QVMAX which borrows the max operator from Q-learning when updating $V(S_t)$:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\Big(R_{t+1} + \gamma V(S_{t+1}) - Q(S_t, A_t)\Big)\,, \tag{5}$$

$$V(S_t) \leftarrow V(S_t) + \alpha\Big(R_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q(S_{t+1}, a') - V(S_t)\Big)\,. \tag{7}$$

One consequence of this change is that now there is reciprocal bootstrapping between Q and V. Previously, with QV-learning, we had only a unidirectional information flow: V bootstrapped from itself, and Q bootstrapped from V. Hence, Q could not corrupt the state values in any way. We illustrate this distinction in Figure 2. Although reciprocal bootstrapping is not necessarily an undesirable property, it is worth noting that a fundamental characteristic of the algorithm has changed.

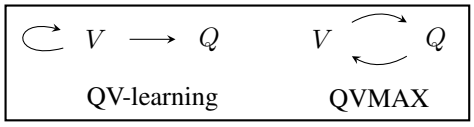

Figure 2: Depiction of bootstrapping in QV-learning variants. The arrows point from the bootstrapped value function to the value function being updated.

Another, more serious consequence of this change is that the update now suffers from off-policy bias. Wiering & van Hasselt (2009) hypothesized that the use of the max operator in Eq. (7) would induce convergence to q_*, just as it does in Q-learning. Unfortunately, this is not true, as we show in the next proposition.

Proposition 3.2. $\left[\begin{smallmatrix} q_* \\ v_* \end{smallmatrix}\right]$ *is generally not a fixed point of QVMAX.*

Proof. See Appendix A.2. □

The update fails to correct the distribution of the observed reward, R_{t+1}, which is collected by the behavior policy. This is because Eq. (7) is *not* conditioned on the state-action pair, (S_t, A_t), as is normally the case for Q-learning, but is instead conditioned only on the state, S_t. The freedom in the choice of A_t means the expected reward observed from state S_t now depends on behavior policy b.

To eliminate the bias, the update cannot depend explicitly on the sampled reward, R_{t+1}. We therefore propose to modify Eq. (7) in the QVMAX updates such that we have

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\Big(R_{t+1} + \gamma V(S_{t+1}) - Q(S_t, A_t)\Big)\,, \tag{5}$$

$$V(S_t) \leftarrow V(S_t) + \alpha\Big(\max_{a \in \mathcal{A}} Q(S_t, a) - V(S_t)\Big)\,. \tag{10}$$

We call this *Bias-Corrected* QVMAX (BC-QVMAX). This is the proper algorithm for learning q_*, as the update now approximates the Bellman optimality relationship $v_*(s) = \max_{a \in \mathcal{A}} q_*(s, a)$.

Proposition 3.3. $\left[\begin{smallmatrix} q_* \\ v_* \end{smallmatrix}\right]$ *is the unique fixed point of BC-QVMAX.*

Proof. See Appendix A.3. □

A peculiarity of the updates in Eqs. (5) and (10) is that they closely resemble Q-learning in Eq. (3). In fact, if we set $\alpha = 1$, then Eq. (10) simply becomes a table overwrite: $V(S_t) \leftarrow \max_{a \in \mathcal{A}} Q(S_t, a)$. This would make the algorithm almost the same as Q-learning, but with a delay introduced before

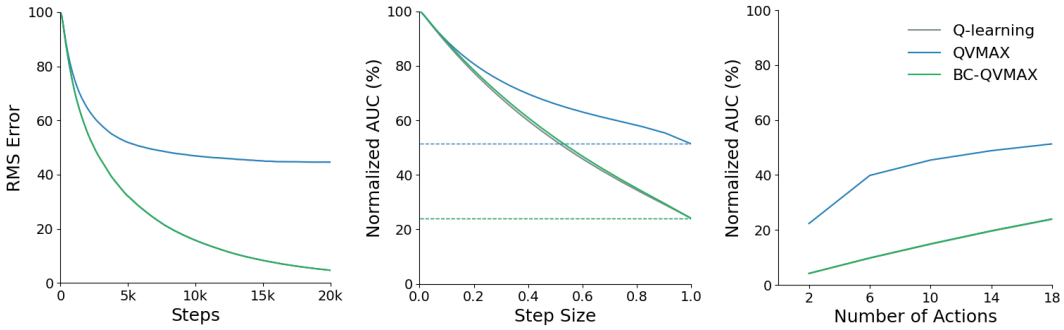

Figure 3: Off-policy control performance of BC-QVMAX compared to QVMAX and Q-learning in the 4-state MDP. Experiment setup is the same as that of Figure 1. (Note that Q-learning is sometimes eclipsed by BC-QVMAX.)

bootstrapping. When $\alpha < 1$, then Eq. (10) has an extra smoothing effect, mixing together stale estimates of the maximum action value in each state and exacerbating this delay.

We slightly modify the prediction experiment from Section 3.1 to demonstrate the bias of QV-MAX. This time, we target a greedy policy with respect to the current action-value function, making Expected Sarsa equivalent to Q-learning. We therefore compare Q-learning, QVMAX, and BC-QVMAX. The optimal policy is to select a_0 unconditionally, making $q_*(s, a_0) = 1 \,/\, (1 - \gamma)$. It follows that $q_*(s, a_i) = \gamma \,/\, (1 - \gamma)$ for $i \neq 0$.

All experiment and plotting procedures remain the same as before, except that now the RMS error is defined in terms of q_*: i.e., $\|q_* - Q\|_2$. We plot the results in Figure 3. As can be seen in Figure 3 (left), and as our theory predicted, QVMAX asymptotes substantially above the zero-error mark due to its biased update. In contrast, BC-QVMAX achieves a much lower error. However, as we also discussed above, the smoothing effect of BC-QVMAX makes it similar to, but rather slower than, Q-learning. The results in Figure 3 (left) are identical only because $\alpha = 1$ happens to be the best step size for both methods, which makes them nearly equivalent (except for the delay mentioned previously). However, for all $\alpha < 1$, we can see from Figure 3 (center) that BC-QVMAX slightly lags behind Q-learning.

These results unfortunately indicate that off-policy control with QV-learning algorithms is much harder than on-policy prediction, since model-free estimation of v_* is not straightforward. While we note that this one experiment is not enough to rule out the viability of QVMAX approaches in this setting, it does provide convincing evidence against it. In particular, the theoretically correct version of the algorithm, BC-QVMAX, is similar to a learned approximation of Q-learning which introduces delay into the bootstrapping. This does not serve an immediately obvious benefit and appears to negatively impact sample efficiency, though there is the possibility that this smoothing effect may have utility in some settings (e.g., to enhance stability in tracking problems).

4 AV-learning Algorithms

Rather than learning explicit state- and action-value functions like QV-learning methods do, an alternative approach is to decompose the action-value function into constituent state-value and advantage functions that can be implicitly updated using gradient descent. This *dueling* strategy was introduced by Wang et al. (2016), but we generalize it in this section.

4.1 Dueling Q-learning

To obtain a tabular Dueling Q-learning update, we evaluate the chain rule in Eq. (9), as discussed earlier in Section 2.2. Q-learning's TD error is $\delta_t = R_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q(S_{t+1}, a') - Q(S_t, A_t)$.

Because $Q(s,a)$ is defined according to Eq. (8), the chain rule preserves the error, δ_t, inside the quadratic term and then scales it in the following manner:

$$\mathrm{Adv}(S_t,a) \leftarrow \mathrm{Adv}(S_t,a) + \alpha\left(\mathbf{1}_{\{a=A_t\}} - \frac{1}{|\mathcal{A}|}\right)\delta_t\,, \quad \forall a \in \mathcal{A}\,,$$
$$V(S_t) \leftarrow V(S_t) + \alpha\delta_t\,.$$

It may seem unusual to write out these updates in this manner; Dueling Q-learning was originally proposed for deep RL, where automatic differentiation would handle the partial derivatives. Nevertheless, these updates are well-defined for tabular value functions. Furthermore, they reveal a property behind this particular style of dueling. At each time step, a single advantage value is incremented in proportion to δ_t, and then the $|\mathcal{A}|$ advantage values are decremented in the same proportion to δ_t / $|\mathcal{A}|$. This implies that the arithmetic mean of the advantages is an *invariant* quantity; if we initialize the advantage values to zero, then the mean will remain zero throughout training.

4.2 Regularized Dueling Q-learning

We propose a new dueling algorithm which does not rely on subtracting the identifiability term in Eq. (8). We start with the most general decomposition of the action-value function,

$$Q(s,a) = V(s) + \mathrm{Adv}(s,a)\,. \qquad (11)$$

This decomposition is unidentifiable (Wang et al., 2016, Sec. 3); we can add a constant to $V(s)$ and subtract the same constant from $\mathrm{Adv}(s,a)$ without changing $Q(s,a)$. Consequently, there are infinitely many valid decompositions that satisfy Eq. (11) and no clear way for an algorithm to choose between them.

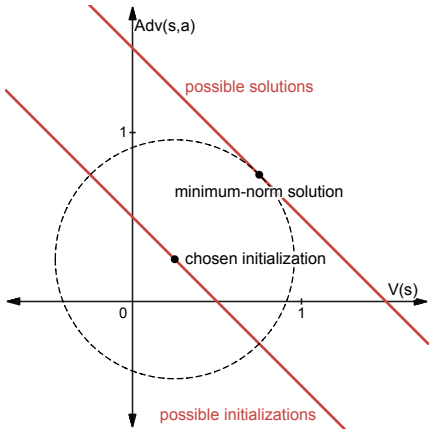

Figure 4: RDQ finds the minimum-norm solution to the underdetermined system of equations $Q(s,a) = V(s) + \mathrm{Adv}(s,a)$, $\forall a \in \mathcal{A}$. Thus, RDQ may require fewer updates than Dueling Q-learning to converge in practice.

Dueling DQN seeks to accurately approximate both $v_\pi(s)$ and $\mathrm{Adv}_\pi(s,a) \stackrel{\text{def}}{=} q_\pi(s,a) - v_\pi(s)$ in order to estimate $q_\pi(s,a)$, which motivates the subtraction of the mean advantage to get Eq. (8) from Eq. (11). The mean advantage is meant to coarsely approximate the expected advantage under the target policy, but sacrifices the precise semantics of v_π and Adv_π. We propose an alternative approach in which we search for *any* two functions, $V(s)$ and $\mathrm{Adv}(s,a)$, that accurately reconstruct $q_\pi(s,a)$, further relaxing these semantics in order to expedite learning.

Let us consider the underdetermined system of equations in Eq. (11) to motivate our algorithm. For a particular state-action pair (s,a), we can represent the specific solution found by Dueling Q-learning as an ordered pair: $(V(s), \mathrm{Adv}(s,a))$. This particular solution is just one of infinitely many ordered pairs that satisfy Eq. (11), which form a negatively sloped line in the V-Adv plane: $\mathrm{Adv}(s,a) = q_\pi(s,a) - V(s)$, depicted in Figure 4. When $Q(s,a)$ is initialized, the locus of valid initializations of $V(s)$ and $\mathrm{Adv}(s,a)$ is also a negatively sloped line: $\mathrm{Adv}(s,a) = Q(s,a) - V(s)$, again depicted in Figure 4. These two lines are parallel, and therefore the shortest directional path between them is the vector orthogonal to both of them—the minimum-norm projection. This fact remains true regardless of how $Q(s,a)$ is initialized, since the slope of both lines is always -1.

For this reason, we hypothesize that a good solution to Eq. (11) is the point on the solution line closest to the initialization of $V(s)$ and $\mathrm{Adv}(s,a)$, since it may require fewer updates to reach. In practice, because the state-value and advantage functions are typically initialized near zero for both tabular and deep RL, we can approximate this solution more simply as the closest point to the origin.

We therefore propose to apply l_2 regularization to Eq. (9) to simultaneously restore identifiability to Eq. (11) and encourage the algorithm to find this minimum-norm approximation. Specifically, we penalize the squared-error loss from Eq. (9) with a term of

$$\frac{1}{2}V(s)^2 + \frac{1}{2}\sum_{a\in\mathcal{A}} \mathrm{Adv}(s,a)^2 \,. \tag{12}$$

Since we now have $Q(s,a) = V(s) + \mathrm{Adv}(s,a)$ from Eq. (11) (no identifiability term), evaluating the chain rule in Eq. (9) gives us the following updates:

$$\mathrm{Adv}(S_t,a) \leftarrow (1-\beta)\,\mathrm{Adv}(S_t,a) + \alpha \mathbf{1}_{\{a=A_t\}}\delta_t \,, \quad \forall a \in \mathcal{A} \,, \tag{13}$$

$$V(S_t) \leftarrow (1-\beta)V(S_t) + \alpha\delta_t \,, \tag{14}$$

where $\beta \in [0,1)$ is the regularization coefficient. We call this new algorithm Regularized Dueling Q-learning (RDQ). Specifically, we refer to this variant with the l_2 penalty as *Soft* RDQ. In practice, we found Soft RDQ to be noisy in the tabular setting, since the state and advantage values become leaky—they get pushed towards zero by the $1-\beta$ factor, and then have to compensate using the stochastic TD error, δ_t. However, Soft RDQ is very useful in the case of function approximation, since the penalty in Eq. (12) is architecture-agnostic and can be easily combined with complex neural networks. We experiment with this in Section 4.3.

In the tabular setting, we do not need an l_2 penalty to obtain the desired minimum-norm solution. This is because differentiating the penalty in Eq. (12) with respect to $V(s)$ and then solving it for zero gives

$$V(s) = \sum_{a\in\mathcal{A}} \mathrm{Adv}(s,a) \,. \tag{15}$$

Furthermore, when we remove the l_2 penalty from Eqs. (13) and (14) by setting $\beta = 0$, we see that both $V(s)$ and exactly one advantage value, $\mathrm{Adv}(s,a)$, are incremented by exactly the same amount at each time step—indicating that Eq. (15) is another invariant quantity. This implies that, regardless of how V and Adv are initialized, Eq. (15) always remains true.[4] Therefore, rather than using an explicit l_2 penalty, we can directly apply the RDQ update without a penalty to achieve the minimum-norm solution in the tabular setting:

$$\mathrm{Adv}(S_t,a) \leftarrow \mathrm{Adv}(S_t,a) + \alpha\,\mathbf{1}_{\{a=A_t\}}\,\delta_t \,, \quad \forall a \in \mathcal{A} \,,$$

$$V(S_t) \leftarrow V(S_t) + \alpha\delta_t \,.$$

We call this variant *Hard* RDQ, since it explicitly follows the minimum-norm path (on average) without the use of a soft penalty. Unfortunately, there does not appear to be a simple extension of this idea to the function approximation case. If we were to apply these update rules under function approximation, the different gradients calculated for $V(s)$ and $\mathrm{Adv}(s,a)$ would violate the invariant quantity in Eq. (15) and lose the minimum-norm convergence property.

To test this new algorithm, we repeat the off-policy control experiment from Section 3.2. We once again use Q-learning as a baseline, and then additionally compare Dueling Q-learning with Hard RDQ. The only minor modifications we make is that we now change the suboptimal reward from 0 to -1 (which does not change the optimal policy), set $\gamma = 0.999$, and initialize $V(s)$ and $\mathrm{Adv}(s,a)$ using zero-mean Gaussian noise with a standard deviation of 2. The rationale for the nonzero reward is so that Q does not correctly estimate either of the actions' true values initially (on average).

We plot the results in Figure 5. All experiment and plotting procedures remain the same as before. In contrast to the QVMAX algorithm tested earlier, both dueling methods dramatically outperform Q-learning, showing that the advantage decomposition is a highly effective strategy in general. In the largest instantiation of our MDP, where $|\mathcal{A}| = 18$, Hard RDQ slightly outperforms Dueling Q-learning across the range of step sizes (see Figure 5; left, center). The spike at $\alpha = 0.5$ is due to

[4] That is, up to a fixed translation determined by the difference at initialization: $V_0(s) - \sum_{a\in\mathcal{A}} \mathrm{Adv}_0(s,a)$.

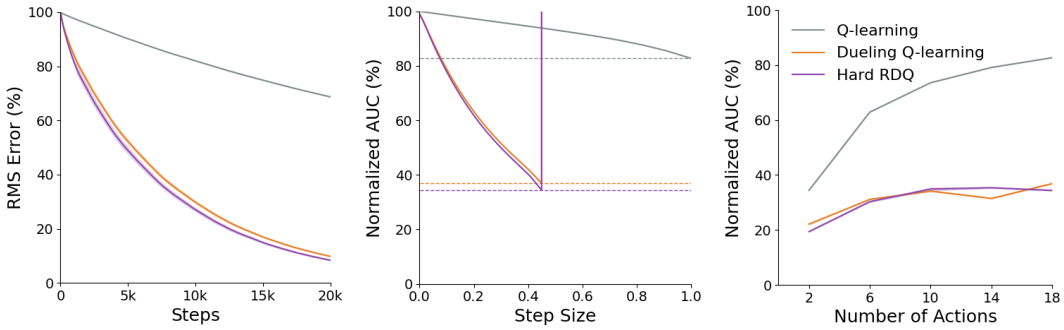

Figure 5: Off-policy control performance of Hard RDQ compared to Dueling Q-learning and Q-learning in the 4-state MDP. Experiment setup is the same as that of Figure 1.

the fact that AV-learning methods sum the updates to V and Adv, making the total effective step size equal to 1 and therefore unstable. Across the various MDP instances (see Figure 5; right), Hard RDQ performs slightly better than Dueling Q-learning in a small majority of cases, but both algorithms perform very well overall.

4.3 Deep RL Experiments

Because Dueling Q-learning was originally proposed as a network architecture for Deep Q-Network (DQN; Mnih et al., 2015), we test the performance of RDQ in a deep RL setting, where a neural network is used to approximate the action-value function. Our benchmark is the MinAtar domain (Young & Tian, 2019), which includes five Atari-like games: Asterix, Breakout, Freeway, Seaquest, Space Invaders. The state representation for MinAtar is 10×10 multi-channel binary images displaying various objects and their velocities. The reward function is the incremental game score.

We compare the soft variant of RDQ against DQN and Dueling DQN, using the hyperparameters used by Obando-Ceron & Castro (2021). The action-value function $Q(s, a; \boldsymbol{\theta}_t)$ is approximated by a neural network, where $\boldsymbol{\theta}_t$ is the network parameters at time t. The network architecture we use for DQN is the same used by Young & Tian (2019): a 16-filter, 3×3 convolutional layer, a 128-unit dense layer, and a final linear layer which maps to the $|\mathcal{A}|$ action values. All layers except the last apply a rectified linear unit (ReLU) activation function. We apply LeCun normal initialization (LeCun et al., 2002) to the network parameters.

For RDQ and Dueling DQN, we create a dueling architecture following Wang et al.'s 2016 procedure. We duplicate the 128-unit hidden layer to branch the network in parallel streams, and then separately map these into linear outputs of size $|\mathcal{A}|$ and 1 for estimating $\mathrm{Adv}(s, a; \boldsymbol{\theta}_t)$ and $V(s, a; \boldsymbol{\theta}_t)$, respectively. These are summed together (with broadcasting) to compute $Q(s, a; \boldsymbol{\theta}_t)$. Dueling DQN additionally subtracts the identifiability term $\frac{1}{|\mathcal{A}|} \sum_{a' \in \mathcal{A}} \mathrm{Adv}(s, a'; \boldsymbol{\theta}_t)$ per Eq. (8).

The agents execute an ϵ-greedy policy, where ϵ is fixed to 1 for the first 1,000 time steps of training and then linearly annealed to 0.01 over the next 250,000 time steps. Each transition, $(S_t, A_t, R_{t+1}, S_{t+1})$, is stored in a replay buffer with a capacity of 100,000 transitions. We let $\mathcal{D}_t$ denote the replay memory at time t. Every 4 time steps, the agents update the parameters using the Adam optimizer (Kingma & Ba, 2015) with a learning rate of 2.5×10^{-4} and a denominator constant of $\epsilon = 3.125 \times 10^{-4}$. Both DQN and Dueling DQN are trained to minimize the loss

$$\mathcal{L}_t^{\mathrm{DQN}} \stackrel{\mathrm{def}}{=} \frac{1}{2} \mathbb{E}\left[\left(R + \gamma \max Q(S, A; \boldsymbol{\theta}_t^-) - Q(S, A; \boldsymbol{\theta}_t)\right)^2\right], \tag{16}$$

where $\boldsymbol{\theta}_t^-$ is the target network parameters copied from $\boldsymbol{\theta}_t$ every 1,000 time steps. The expectation in Eq. (16) is taken over the uniform distribution of samples (S, A, R, S') in $\mathcal{D}_t$; in practice, we approx-

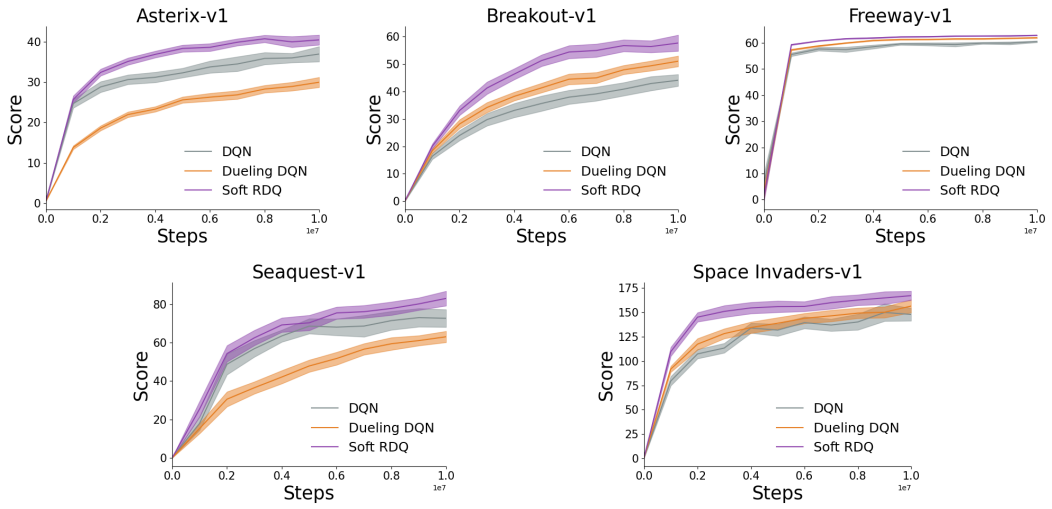

Figure 6: Deep RL results across five MinAtar games. Each algorithm is run for 30 seeds in each environment. The dark lines represent the mean across these 30 seeds, and the shaded region corresponds to a 95% confidence interval.

imate this with minibatches of size 32. RDQ inherits the same loss, but adds the regularization term:

$$\mathcal{L}_t^{\text{RDQ}} \stackrel{\text{def}}{=} \mathcal{L}_t^{\text{DQN}} + \frac{\beta}{2}\,\mathbb{E}\left[V(S;\boldsymbol{\theta}_t)^2 + \sum_{a\in\mathcal{A}} \text{Adv}(S,a;\boldsymbol{\theta}_t)^2\right] .$$

We choose $\beta = 10^{-3}$ for the regularization strength; we did not tune this value.

Each agent was trained for a total of 10 million time steps. Every 1 million time steps, the agent was evaluated with an ϵ-greedy policy for 1,000 episodes with $\epsilon = 0.01$. In Figure 6, we plot the mean undiscounted return for the evaluation episodes as a function of training time; these results are averaged over 30 independent trials and the shading indicates 95% confidence intervals.

Figure 6 depicts our results. We see that between DQN and Dueling DQN, there does not appear to be a clearly superior algorithm, and one algorithm may outperform the other depending on the environment. However, we also see that RDQ significantly outperforms DQN and Dueling DQN in all five environments. Note that RDQ shares the same architecture as Dueling DQN. Although $\beta = 10^{-3}$ works well in this setting, more experiments would be needed to determine its efficacy in other domains. Given that we did not tune β, it is also possible that performance could improve if tuned.

5 Conclusion

In this paper, we made several advances in the understanding of TD-learning algorithms in which state-value functions are learned in tandem with action-value functions. We showed that on-policy QV-learning algorithms have benefits for prediction, as in the original setting for which they were proposed, but tend to be much slower when the target policy is greedy. We also demonstrated that the prevailing off-policy control variant of QV-learning, QVMAX, is biased, and we introduced a new variant called BC-QVMAX which empirically restores convergence. Lastly, we introduced a novel Dueling Q-learning method called RDQ, which does not require the subtraction of an identifiability term, and has desirable properties in terms of the geometry of the learned value functions. In a deep RL setting based on the MinAtar games, we showed that RDQ greatly outperforms Dueling DQN, despite having identical network architectures and the same, well-tuned hyperparameters. Although there is still much to learn about these algorithms, our analysis helps to clarify their efficacy and distinguishing properties, and has already demonstrated potential for the development of new and effective RL algorithms.

Acknowledgments

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada CIFAR AI Chair Program. Prabhat Nagarajan has been additionally supported by the Alberta Innovates Graduate Student Scholarship. Computational resources were provided in part by the Digital Research Alliance of Canada.

References

Leemon C. Baird. Advantage Updating. Technical Report WL-TR-93-1146, Wright-Patterson Air Force Base, 1993.

Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3(1):133–181, 1922.

Richard Bellman. *Dynamic Programming*. Princeton University Press, 1957.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

Brett Daley. *Multistep Credit Assignment in Deep Reinforcement Learning*. PhD thesis, University of Alberta, 2025.

George H. John. When the best move isn't optimal: Q-learning with exploration. In *AAAI Conference on Artificial Intelligence (AAAI)*, 1994.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Yann LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. Efficient BackProp. In *Neural Networks: Tricks of the Trade*. Springer, 2002.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Joseph Modayil and Zaheer Abbas. Towards model-free RL algorithms that scale well with unstructured data, 2023. arXiv:2311.02215.

Johan Samir Obando-Ceron and Pablo Samuel Castro. Revisiting Rainbow: Promoting more insightful and inclusive deep reinforcement learning research. In *International Conference on Machine Learning (ICML)*, 2021.

Gavin A. Rummery and Mahesan Niranjan. On-line Q-Learning using connectionist systems. Technical Report CUED/F-INFENG/TR 166, University of Cambridge, 1994.

Matthia Sabatelli, Gilles Louppe, Pierre Geurts, and Marco A. Wiering. The deep quality-value family of deep reinforcement learning algorithms. In *International Joint Conference on Neural Networks (IJCNN)*, 2020.

Satinder Singh, Tommi Jaakkola, Michael L. Littman, and Csaba Szepesvári. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38:287–308, 2000.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

Yunhao Tang, Rémi Munos, Mark Rowland, and Michal Valko. VA-learning as a more efficient alternative to Q-learning. In *International Conference on Machine Learning (ICML)*, 2023.

Hado van Hasselt. Double Q-learning. In *Neural Information Processing Systems (NeurIPS)*, 2010.

Hado van Hasselt. *Insights in Reinforcement Learning*. PhD thesis, Utrecht University, 2011.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2016.

Christopher J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, University of Cambridge, 1989.

Marco A. Wiering. QV(λ)-learning: A new on-policy reinforcement learning algorithm. In *European Workshop on Reinforcement Learning (EWRL)*, 2005.

Marco A. Wiering and Hado van Hasselt. Two novel on-policy reinforcement learning algorithms based on TD(λ)-methods. In *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, 2007.

Marco A. Wiering and Hado van Hasselt. The QV family compared to other reinforcement learning algorithms. In *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, 2009.

Kenny Young and Tian Tian. MinAtar: An atari-inspired testbed for thorough and reproducible reinforcement learning experiments, 2019. arXiv:1903.03176.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Proofs

This section contains the proofs for all theoretical results in the paper. These results rely on the analysis of dynamic-programming operators for MDPs, which we introduce here rather than Section 2 for clarity of exposition.

In this context, value functions can be represented as vectors: $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{S}|}$ for state values and $\boldsymbol{q} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ for action values. Each element of a vector corresponds to the value estimate for a state or a state-action pair; the order of the elements does not matter as long as it is consistent. Likewise, the expected reward

$$r(s,a) \stackrel{\text{def}}{=} \sum_{r\in\mathcal{R}} r \sum_{s'\in\mathcal{S}} p(s', r \mid s, a)$$

is an action-value function which can be represented as a vector, $\boldsymbol{r} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$.

Operators are then mappings between these vector spaces, which often represent an expected TD update applied simultaneously to every element. In other words, TD methods are asynchronous, stochastic approximations of their underlying dynamic-programming operators. Analyzing these operators gives insight into convergence properties.

We adopt the operator convention of Daley (2025), summarized in Table 1, which decomposes the standard Bellman operators into three fundamental operators: one for state transitions ($\boldsymbol{P}$) and two for action selection ($\boldsymbol{E}_\pi$ and E). The major benefit to this notation is that these fundamental operators are never overloaded; each applies to either $\boldsymbol{v}$ or $\boldsymbol{q}$ but not both. This allows us to expand the Bellman operators for $\boldsymbol{v}$ and $\boldsymbol{q}$ into unambiguous expressions below—especially useful for analyzing QV-learning methods.

Given these preliminaries, the Bellman operator, T_π, is overloaded such that

$$T_\pi \boldsymbol{q} \stackrel{\text{def}}{=} \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{E}_\pi \boldsymbol{q}\,,$$
$$T_\pi \boldsymbol{v} \stackrel{\text{def}}{=} \boldsymbol{E}_\pi(\boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{v})\,.$$

The Bellman expectation equations, Eqs. (1) and (4), can now be expressed succinctly as $\boldsymbol{q}_\pi = T_\pi \boldsymbol{q}_\pi$ and $\boldsymbol{v}_\pi = T_\pi \boldsymbol{v}_\pi$. In other words, the unique fixed point of T_π is either $\boldsymbol{q}_\pi$ or $\boldsymbol{v}_\pi$ depending on whether it acts on action values or state values, respectively. Additionally, these fixed points are related by $\boldsymbol{v}_\pi = \boldsymbol{E}_\pi \boldsymbol{q}_\pi$.

Table 1: Fundamental operators for MDPs (Daley, 2025, Ch. 2).

Symbol	Input/Output	Definition
$\boldsymbol{P}$	$\mathbb{R}^{\lvert\mathcal{S}\rvert} \to \mathbb{R}^{\lvert\mathcal{S}\times\mathcal{A}\rvert}$	$(Pv)(s,a) \stackrel{\text{def}}{=} \sum_{s'\in\mathcal{S}} v(s') \sum_{r\in\mathcal{R}} p(s', r \mid s, a)$
$\boldsymbol{E}_\pi$	$\mathbb{R}^{\lvert\mathcal{S}\times\mathcal{A}\rvert} \to \mathbb{R}^{\lvert\mathcal{S}\rvert}$	$(E_\pi q)(s) \stackrel{\text{def}}{=} \sum_{a\in\mathcal{A}} \pi(a \mid s)\, q(s,a)$
E	$\mathbb{R}^{\lvert\mathcal{S}\times\mathcal{A}\rvert} \to \mathbb{R}^{\lvert\mathcal{S}\rvert}$	$(Eq)(s) \stackrel{\text{def}}{=} \max_{a\in\mathcal{A}} q(s,a)$

The Bellman optimality operator, T, is overloaded similarly but uses the greedy-policy operator $\boldsymbol{E}$ instead of the expected-policy operator $\boldsymbol{E}_\pi$:

$$T\boldsymbol{q} \stackrel{\text{def}}{=} \boldsymbol{r} + \gamma \boldsymbol{P} E \boldsymbol{q}\,,$$
$$T\boldsymbol{v} \stackrel{\text{def}}{=} E(\boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{v})\,.$$

The subtle difference here is that T is a nonlinear operator and admits the optimal value functions as the fixed points: $\boldsymbol{q}_* = T\boldsymbol{q}_*$ and $\boldsymbol{v}_* = T\boldsymbol{v}_*$. These fixed points are analogously related by $\boldsymbol{v}_* = E\boldsymbol{q}_*$.

These complete our operator definitions. Our following theoretical results in the remainder of this section are derived from the fundamental properties of these operators.

A.1 Proof of Theorem 3.1

Theorem 3.1 (QV-learning Contraction)**.** *The expected QV-learning update corresponds to an affine joint operator* $H\colon \left[\begin{smallmatrix}\boldsymbol{q}\\ \boldsymbol{v}\end{smallmatrix}\right] \mapsto \boldsymbol{b} + \boldsymbol{A}\left[\begin{smallmatrix}\boldsymbol{q}\\ \boldsymbol{v}\end{smallmatrix}\right]$*, where* $\boldsymbol{b} = \left[\begin{smallmatrix}\boldsymbol{r}\\ \boldsymbol{E}_b\boldsymbol{r}\end{smallmatrix}\right]$ *and* $\boldsymbol{A} = \gamma\left[\begin{smallmatrix}\boldsymbol{0} & \boldsymbol{P}\\ \boldsymbol{0} & \boldsymbol{E}_b\boldsymbol{P}\end{smallmatrix}\right]$*. The operator* H *is a contraction mapping with its unique fixed point equal to* $\left[\begin{smallmatrix}\boldsymbol{q}_b\\ \boldsymbol{v}_b\end{smallmatrix}\right]$*.*

Proof. The operator updates corresponding to the QV-learning, Eqs. (5) and (6), are

$$\boldsymbol{q} = \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{v}\,,$$
$$\boldsymbol{v} = \boldsymbol{E}_b(\boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{v})\,.$$

The fixed points of these updates are individually $\boldsymbol{q}_b$ and $\boldsymbol{v}_b$, respectively, which we show now by substituting these quantities. The first update returns $\boldsymbol{q}_b$ because of the relation $\boldsymbol{v}_b = \boldsymbol{E}_b\boldsymbol{q}_b$:

$$\boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{v}_b = \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{E}_b\boldsymbol{q}_b = T_b\boldsymbol{q}_b = \boldsymbol{q}_b\,.$$

The second update is just equivalent to the Bellman operator, T_b, applied to $\boldsymbol{v}$, which admits $\boldsymbol{v}_b$ as its unique fixed point.

We next write these updates as a joint operator update

$$\begin{bmatrix}\boldsymbol{q}\\ \boldsymbol{v}\end{bmatrix} \leftarrow \underbrace{\begin{bmatrix}\boldsymbol{r}\\ \boldsymbol{E}_b\boldsymbol{r}\end{bmatrix}}_{\boldsymbol{b}} + \underbrace{\gamma\begin{bmatrix}\boldsymbol{0} & \boldsymbol{P}\\ \boldsymbol{0} & \boldsymbol{E}_b\boldsymbol{P}\end{bmatrix}}_{\boldsymbol{A}}\begin{bmatrix}\boldsymbol{q}\\ \boldsymbol{v}\end{bmatrix}\,,$$

which we just established has a fixed point at $\left[\begin{smallmatrix}\boldsymbol{q}_b\\ \boldsymbol{v}_b\end{smallmatrix}\right]$. Moreover, this is an affine operator of the form $H\colon \boldsymbol{y} \mapsto \boldsymbol{b} + \boldsymbol{A}\boldsymbol{y}$, where we let $\boldsymbol{y} = \left[\begin{smallmatrix}\boldsymbol{q}\\ \boldsymbol{v}\end{smallmatrix}\right]$ for brevity. To complete the proof, we show that H is a maximum-norm contraction mapping, which implies the fixed point $\left[\begin{smallmatrix}\boldsymbol{q}_b\\ \boldsymbol{v}_b\end{smallmatrix}\right]$ is unique. Because each row of $\boldsymbol{A}$ is a probability distribution scaled by γ, it follows that $\|\boldsymbol{A}\|_\infty = \gamma$. For any two vectors $\boldsymbol{y}, \boldsymbol{y}' \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|+|\mathcal{S}|}$, we therefore have

$$\|H\boldsymbol{y} - H\boldsymbol{y}'\|_\infty = \|\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{y}')\|_\infty \le \gamma\,\|\boldsymbol{y} - \boldsymbol{y}'\|_\infty\,,$$

so H is indeed a maximum-norm contraction mapping. By the Banach fixed-point theorem (Banach, 1922), the fixed point $\left[\begin{smallmatrix}\boldsymbol{q}_b\\ \boldsymbol{v}_b\end{smallmatrix}\right]$ is unique and hence $\lim_{i\to\infty} H^i\boldsymbol{y} = \left[\begin{smallmatrix}\boldsymbol{q}_b\\ \boldsymbol{v}_b\end{smallmatrix}\right]$ for any initial vector $\boldsymbol{y}$. □

A.2 Proof of Proposition 3.2

Proposition 3.2. $\left[\begin{smallmatrix}\boldsymbol{q}_*\\ \boldsymbol{v}_*\end{smallmatrix}\right]$ *is generally not a fixed point of QVMAX.*

Proof. The operator updates corresponding to Eqs. (5) and (7) are

$$\boldsymbol{q} \leftarrow \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{v}\,,$$
$$\boldsymbol{v} \leftarrow \boldsymbol{E}_b(\boldsymbol{r} + \gamma \boldsymbol{P}E\boldsymbol{q})\,.$$

If $\left[\begin{smallmatrix}\boldsymbol{q}_*\\ \boldsymbol{v}_*\end{smallmatrix}\right]$ were a fixed point of QVMAX, then substituting these vectors into the above operator updates would return the same result.

The first update (which reassigns $\boldsymbol{q}$) correctly remains invariant when applied to the fixed point. This is because $\boldsymbol{v}_* = E\boldsymbol{q}_*$ and therefore

$$\begin{aligned}\boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{v}_* &= \boldsymbol{r} + \gamma \boldsymbol{P} E \boldsymbol{q}_* \\ &= T\boldsymbol{q}_* \\ &= \boldsymbol{q}_* \,.\end{aligned}$$

However, the second update (which reassigns $\boldsymbol{v}$) does not remain invariant when applied to the fixed point—the root cause of the bias in QVMAX. This can be seen because

$$\begin{aligned}\boldsymbol{E}_b(\boldsymbol{r} + \gamma \boldsymbol{P} E \boldsymbol{q}_*) &= \boldsymbol{E}_b(\boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{v}_*) \\ &\neq E(\boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{v}_*) \\ &= T\boldsymbol{v}_* \\ &= \boldsymbol{v}_* \,.\end{aligned}$$

This mismatch stems from the fact that the updates to $\boldsymbol{v}$ are conditioned only on the states of the MDP and do not correct for the influence of taken actions. As such, the updates are subject to changes to the reward and next-state distributions induced by the particular behavior policy, b. This manifests as the discrepancy $\boldsymbol{E}_b \neq E$ in the above equations. These two operators only coincide in the serendipitous case that b happens to be greedy with respect to the expected 1-step returns, $\boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{v}$, but not in general. Ultimately, the above analysis demonstrates that $\left[\begin{smallmatrix}\boldsymbol{q}_*\\ \boldsymbol{v}_*\end{smallmatrix}\right]$ cannot be the fixed point of QVMAX for an arbitrary behavior policy b and MDP, which completes the proof. □

A.3 Proof of Proposition 3.3

Proposition 3.3. $\left[\begin{smallmatrix}\boldsymbol{q}_*\\ \boldsymbol{v}_*\end{smallmatrix}\right]$ *is the unique fixed point of BC-QVMAX.*

Proof. The operator updates corresponding to BC-QVMAX, Eqs. (5) and (10), are

$$\begin{aligned}\boldsymbol{q} &\leftarrow \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{v} \,, \\ \boldsymbol{v} &\leftarrow E\boldsymbol{q} \,.\end{aligned}$$

We can unroll these updates by substituting them into each other—exploiting the fact that the updates are interleaved and not conducted simultaneously.

The first update (which reassigns $\boldsymbol{q}$) becomes

$$\begin{aligned}\boldsymbol{q} &\leftarrow \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{v} \\ &= \boldsymbol{r} + \gamma \boldsymbol{P} E \boldsymbol{q} \\ &= T\boldsymbol{q} \,.\end{aligned}$$

The second update (which reassigns $\boldsymbol{v}$) becomes

$$\begin{aligned}\boldsymbol{v} &\leftarrow E\boldsymbol{q} \\ &= E(\boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{v}) \\ &= T\boldsymbol{v} \,.\end{aligned}$$

Both unrolled updates are equivalent to their overloaded Bellman optimality operators, T, which are contraction mappings and respectively admit $\boldsymbol{q}_*$ and $\boldsymbol{v}_*$ as their unique fixed points. Therefore, $\left[\begin{smallmatrix}\boldsymbol{q}_*\\ \boldsymbol{v}_*\end{smallmatrix}\right]$ is the unique fixed point of BC-QVMAX. □