Differential Eligibility Vectors for Advantage Updating and Gradient Methods

Francisco S. Melo
Instituto Superior Técnico/INESC-ID
TagusPark, Edifício IST
2780-990 Porto Salvo, Portugal
e-mail: fmelo@inesc-id.pt

Abstract

In this paper we propose *differential eligibility vectors* (DEV) for temporal-difference (TD) learning, a new class of eligibility vectors designed to bring out the contribution of each action in the TD-error at each state. Specifically, we use DEV in TD-$Q(\lambda)$ to more accurately learn the relative value of the actions, rather than their absolute value. We identify conditions that ensure convergence w.p.1 of TD-$Q(\lambda)$ with DEV and show that this algorithm can also be used to directly approximate the *advantage function* associated with a given policy, without the need to compute an auxiliary function – something that, to the extent of our knowledge, was not known possible. Finally, we discuss the integration of DEV in LSTDQ and actor-critic algorithms.

1 Introduction

In the reinforcement learning literature it is possible to identify two major classes of methods to address stochastic optimal control problems. The first class comprises *value-based algorithms*, in which the optimal policy is derived from a *value-function*, the latter being the focus of the learning algorithm (Antos, Szepesvári, and Munos 2008; Boyan 2002; Perkins and Precup 2003; Melo, Meyn, and Ribeiro 2008). The second class comprises *policy-based methods*, in which the optimal policy is computed by direct optimization in policy space (Baxter and Bartlett 2001; Sutton et al. 2000; Marbach and Tsitsiklis 2001).[1] Unfortunately, value-based methods typically approximate the target value-function *in average* (Tsitsiklis and Van Roy 1996; Szepesvári and Smart 2004), with no specific concern on how suitable the obtained approximation is in the action selection process (Kakade and Langford 2002).[2] Policy-based methods, on the other hand, typically exhibit large variance and can exhibit prohibitively long learning times (Konda and Tsitsiklis 2003; Kakade and Langford 2002).

In this paper we address the aforementioned drawback of value-based methods. We adapt an existing algorithm—namely TD-Q—making it more suited for control scenarios. In particular, we introduce *differential eligibility vectors* (DEV) as a way to modify TD-$Q(\lambda)$ to more finely discriminate differences in value between the different actions, making it more adequate for action selection in control settings. We further show that this modified version of TD-Q can be used to directly compute an approximation of the *advantage function* (Baird 1993) without the need to explicitly compute a separate value function, which, to the extent of our knowledge, was not known possible until now. Finally, we discuss the application of DEV in a batch RL algorithm (LSTDQ) and the application of our results in a policy gradient setting, further bridging value and policy-based methods.

2 Background

A *Markov decision problem* (MDP) is a tuple $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathsf{P}, r, \gamma)$, where $\mathcal{X} \subset \mathbb{R}^p$ is the compact set of possible states, $\mathcal{A}$ is the finite set of possible actions, $\mathsf{P}_a(x, U)$ represents the probability of moving from state $x \in \mathcal{X}$ to the (measurable) set $U \subset \mathcal{X}$ by choosing action $a \in \mathcal{A}$ and $r(x, a)$ denotes the immediate reward received for taking action a in state x.[3] The constant γ is a discount factor such that $0 \leq \gamma < 1$.

A stationary Markov policy is any mapping π defining for each $x \in \mathcal{X}$ a probability distribution $\pi(x, \cdot)$ over $\mathcal{A}$. Any fixed policy π thus induces a (non-controlled) Markov chain $(\mathcal{X}, \mathsf{P}_\pi)$ where

$$\begin{aligned}\mathsf{P}_\pi(x, U) &\triangleq \mathbb{P}\left[X_{t+1} \in U \mid X_t = x\right] \\ &= \sum_a \pi_t(x, a)\mathsf{P}_a(x, U), \quad \text{for all } t.\end{aligned}$$

$\mathsf{V}^\pi(x)$ denotes the expected sum of discounted rewards obtained by starting at state x and following policy π thereafter,

$$\begin{aligned}&\mathsf{V}^\pi(x) \\ &\triangleq \mathbb{E}_{A_t \sim \pi(X_t), X_{t+1} \sim \mathsf{P}_\pi(X_t)}\left[\sum_{t=0}^{\infty} \gamma^t r(X_t, A_t) \mid X_0 = x\right],\end{aligned}$$

[1] Interestingly, the celebrated *policy-gradient theorem* (Marbach and Tsitsiklis 2001; Sutton et al. 2000) provides an important bridge between the two classes of methods, establishing that policy gradients depend critically on the value functions estimated by value-based methods.

[2] We refer to Example 1 for a small illustration of this drawback.

[3] In the remainder of the paper, we assume $r(x, a)$ is bounded in absolute value by some constant $K > 0$.

where $A_t \sim \pi(X_t)$ means that A_t is drawn according to the distribution $\pi(X_t, \cdot)$ for all t and $X_{t+1} \sim \mathsf{P}_\pi(X_t)$ means that X_{t+1} is drawn according to the transition probabilities associated with the Markov chain defined by policy π. We define the function $\mathsf{Q}^\pi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ as

$$\mathsf{Q}^\pi(x, a) = \mathbb{E}_{Y \sim \mathsf{P}_a(x)}\left[r(x, a) + \gamma \mathsf{V}^\pi(Y)\right], \tag{1}$$

and the *advantage function* associated with π as $\mathsf{A}^\pi(x, a) = \mathsf{Q}^\pi(x, a) - \mathsf{V}^\pi(x)$ (Baird 1993). A policy π^* is *optimal* if, for every $x \in \mathcal{X}$, $\mathsf{V}^{\pi^*}(x) \geq \mathsf{V}^\pi(x)$ for any policy π. We denote by V^* the value-function associated with an optimal policy and by Q^* the corresponding Q-function. The functions V^π and Q^π are known to verify

$$\mathsf{V}^\pi(x) = \mathbb{E}_{A \sim \pi(x), Y \sim \mathsf{P}_A(x)}\left[r(x, A) + \gamma \mathsf{V}^\pi(Y)\right]$$
$$\mathsf{Q}^\pi(x, a) = \mathbb{E}_{Y \sim \mathsf{P}_a(x), A' \sim \pi(Y)}\left[r(x, a) + \gamma \mathsf{Q}^\pi(Y, A')\right],$$

where the expectation with respect to (w.r.t.) A, A' is replaced by a maximization over actions in the case of the optimal functions.

Temporal Difference Learning

Let $\mathcal{V}$ be a parameterized linear family of real-valued functions. A function in $\mathcal{V}$ is any mapping $\mathsf{V} : \mathcal{X} \times \mathbb{R}^M \to \mathbb{R}$ such that $\mathsf{V}(x, \boldsymbol{\theta}) = \sum_{i=1}^M \phi_i(x)\theta_i = \boldsymbol{\phi}^\top(x)\boldsymbol{\theta}$, where the functions $\phi_i : \mathcal{X} \to \mathbb{R}$, $i = 1, \ldots, M$, form a basis for the linear space $\mathcal{V}$, θ_i denotes the ith component of the parameter vector $\boldsymbol{\theta}$ and $^\top$ denotes the transpose operator.

For an MDP $(\mathcal{X}, \mathcal{A}, \mathsf{P}, r, \gamma)$, let $\{x_t\}$ be an infinite sampled trajectory of the chain $(\mathcal{X}, \mathsf{P}_\pi)$, where π is a policy whose value function, V^π, is to be evaluated. Let $\{a_t\}$ denote the corresponding action sequence and $\hat{\mathsf{V}}(\boldsymbol{\theta}_t)$ the estimate of V^π at time t. The *temporal difference* at time t is

$$\delta_t \triangleq r(x_t, a_t) + \gamma \hat{\mathsf{V}}(x_{t+1}, \boldsymbol{\theta}_t) - \hat{\mathsf{V}}(x_t, \boldsymbol{\theta}_t)$$

and can be interpreted as a sample of a one-time step prediction error, *i.e.*, the error between the current estimate of the value-function at state x_t, $\hat{\mathsf{V}}(x_t, \boldsymbol{\theta}_t)$, and a one step-ahead "corrected" estimate, $r(x_t, a_t) + \gamma \hat{\mathsf{V}}(x_{t+1}, \boldsymbol{\theta}_t)$. In its most general formulation, TD(λ) is defined by the update rule

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \alpha_t \delta_t \mathbf{z}_t \qquad \mathbf{z}_{t+1} \leftarrow \gamma\lambda \mathbf{z}_t + \boldsymbol{\phi}(x_{t+1}),$$

where λ is a constant such that $0 \leq \lambda \leq 1$ and $\mathbf{z}_0 = 0$. The vectors $\mathbf{z}_t$ are known as *eligibility vectors* and essentially keep track of how the information from the current sample "corrects" previous updates of the parameter vector.

Policy Gradient

Let $\pi_{\boldsymbol{\omega}}$ be a stationary policy parameterized by some finite-dimensional vector $\boldsymbol{\omega} \in \mathbb{R}^N$. Assume, in particular, that $\pi_{\boldsymbol{\omega}}$ is continuously differentiable w.r.t. $\boldsymbol{\omega}$. Given some probability measure μ over $\mathcal{X}$, we define $\rho(\pi_{\boldsymbol{\omega}}) = (1 - \gamma)\mathbb{E}_{X \sim \mu}\left[\mathsf{V}^{\pi_{\boldsymbol{\omega}}}(X)\right]$. We abusively write $\rho(\boldsymbol{\omega})$ instead of $\rho(\pi_{\boldsymbol{\omega}})$ to simplify the notation. Let $\mathsf{K}^\gamma_{\boldsymbol{\omega}}$ denote the γ-resolvent associated with the Markov chain induced by $\pi_{\boldsymbol{\omega}}$ (Meyn and Tweedie 1993) and $\mu^\gamma_{\boldsymbol{\omega}}$ denote the measure defined as $\mu^\gamma_{\boldsymbol{\omega}}(U) = \mathbb{E}_{X \sim \mu}\left[\mathsf{K}^\gamma_{\boldsymbol{\omega}}(X, U)\right]$, where U is some measurable set $U \subset \mathcal{X}$. Let $\{\psi_i, i = 1, \ldots, M\}$ be a set of linearly independent functions, with $\psi_i : \mathcal{X} \times \mathcal{A} \to \mathbb{R}, i = 1, \ldots, M$, and let $\mathcal{G}$ denote its linear span. Let $\Pi_{\mathcal{G}}$ denote the orthogonal projection onto $\mathcal{G}$ w.r.t. the inner product

$$\langle f, g \rangle_{\boldsymbol{\omega}} = \mathbb{E}_{X \sim \mu^\gamma_{\boldsymbol{\omega}}}\left[\sum_{a \in \mathcal{A}} \pi(X, a) f(X, a) g(X, a)\right].$$

We wish to compute the parameter vector $\boldsymbol{\omega}^*$ such that the corresponding policy $\pi_{\boldsymbol{\omega}^*}$ maximizes ρ. If ρ is differentiable w.r.t. $\boldsymbol{\omega}$, this can be achieved by updating $\boldsymbol{\omega}$ according to

$$\boldsymbol{\omega}_{t+1} \leftarrow \boldsymbol{\omega}_t + \beta_t \nabla_{\boldsymbol{\omega}} \rho(\boldsymbol{\omega}_t),$$

where $\{\beta_t\}$ is a step-size sequence and $\nabla_{\boldsymbol{\omega}}$ denotes the gradient w.r.t. $\boldsymbol{\omega}$. If

$$\boldsymbol{\psi}(x, a) = \nabla_{\boldsymbol{\omega}} \log(\pi_{\boldsymbol{\omega}}(x, a)) \tag{2}$$

then it has been shown that $\nabla_\omega \rho(\omega) = \langle \boldsymbol{\psi}, \Pi_{\mathcal{G}} \mathsf{Q}^{\pi_{\boldsymbol{\omega}}} \rangle$ (Sutton et al. 2000; Marbach and Tsitsiklis 2001). We recall that basis functions verifying (2) are usually referred as *compatible basis functions*. It is also worth noting that in the gradient expression above we can add an arbitrary *baseline function* $b(x)$ to $\Pi_{\mathcal{G}} \mathsf{Q}^{\pi_{\boldsymbol{\omega}}}$ without affecting the gradient, since $\sum_{a \in \mathcal{A}} \nabla_{\boldsymbol{\omega}} \pi_{\boldsymbol{\omega}}(x, a) b(x) = 0$. If b is chosen so as to minimize the variance of the estimate of $\nabla_\omega \rho(\omega)$, the optimal choice of baseline function is $b(x) = \mathsf{V}^{\pi_{\boldsymbol{\omega}}}(x)$ (Bhatnagar et al. 2007). Recalling that the advantage function associated with a policy π is defined as $\mathsf{A}^\pi(x, a) = \mathsf{Q}^\pi(x, a) - \mathsf{V}^\pi(x)$, we get that $\nabla_\omega \rho(\omega) = \langle \boldsymbol{\psi}, \Pi_{\mathcal{G}} \mathsf{A}^{\pi_{\boldsymbol{\omega}}} \rangle$. Finally, by using a *natural gradient* instead of a vanilla gradient (Kakade 2001), an appropriate choice of metric in policy space leads to a further simplification of the above expression, yielding $\tilde{\nabla}_{\boldsymbol{\omega}} \rho(\boldsymbol{\omega}) = \boldsymbol{\theta}^*$, where $\tilde{\nabla}_{\boldsymbol{\omega}}$ is the natural gradient of ρ w.r.t. $\boldsymbol{\omega}$ and $\boldsymbol{\theta}^*$ is such that

$$\boldsymbol{\psi}^\top(x, a)\boldsymbol{\theta}^* = \Pi_{\mathcal{G}} \mathsf{A}^{\pi_{\boldsymbol{\omega}}}(x, a). \tag{3}$$

3 TD-Learning for Control

In this section we discuss some limitations of TD-learning if used to estimate Q^π in a control setting. We then introduce *differential eligibility vectors* to tackle this limitation.

Differential Eligibility Vectors (DEV)

Given a fixed policy π, the TD(λ) algorithm described in Section 2 can be trivially modified to compute an approximation of Q^π instead of V^π, which can then be used to perform greedy policy updates, in a process known as *approximate policy iteration* (Perkins and Precup 2003; Melo, Meyn, and Ribeiro 2008). This "modified" TD(λ), henceforth referred as TD-Q, can easily be described by considering a parameterized linear family $\mathcal{Q}$ of real-valued functions $\mathsf{Q}(\boldsymbol{\theta}) : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$. The TD-$Q$ update is

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \alpha_t \delta_t \mathbf{z}_t \tag{4}$$
$$\mathbf{z}_{t+1} \leftarrow \gamma\lambda \mathbf{z}_t + \boldsymbol{\phi}(x_{t+1}, a_{t+1}), \tag{5}$$

where now

$$\delta_t \triangleq r(x_t, a_t) + \gamma \hat{\mathsf{Q}}(x_{t+1}, a_{t+1}, \boldsymbol{\theta}_t) - \hat{\mathsf{Q}}(x_t, a_t, \boldsymbol{\theta}_t).$$

One drawback of TD-$Q(\lambda)$ is that it seeks to minimize the *overall error* in estimating Q^π, to some extent ignoring the distinction between actions in the MDP. We propose the use of *differential eligibility vectors* that seek to bring out the distinctions between different actions in terms of corresponding Q-values. The approximation computed by TD-Q with DEV is potentially more adequate in control settings.

Let us start by considering the TD-Q update for the simpler case in which $\lambda = 0$:

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \alpha_t \delta_t \boldsymbol{\phi}(x_t, a_t) \tag{6}$$

We can interpret the eligibility vector – in this case $\boldsymbol{\phi}(x_t, a_t)$ – as "distributing" the "error" δ_t among the different components of $\boldsymbol{\theta}$, proportionally to their contribution for this error. However, for the purpose of policy optimization, it would be convenient to differentiate the contribution of the different *actions* to this error. To see why this is so, consider the following extended version of the example above.

Example 1. Let $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathsf{P}, r, \gamma)$ be a single-state MDP, with action-space $\mathcal{A} = \{a, b\}$, $r = [5, 1]$ and $\gamma = 0.95$. We want to compute Q^π for the policy $\pi = [0.5, 0.5]$ and use two basis functions $\phi_1 = [1, 2]$ and $\phi_2 = [1, 1]$. The parameter vector is initialized as $\boldsymbol{\theta} = [0, 0]^\top$ and, for simplicity, we consider $\alpha_t \equiv 1$ in the learning algorithm. Notice that, for the given policy, $\mathsf{Q}^\pi = [62, 58]$, which can be represented exactly by our basis functions by taking $\boldsymbol{\theta}^* = [-4, 66]^\top$. Suppose that $A_0 = a$. We have

$$\boldsymbol{\theta}_1(1) \leftarrow \boldsymbol{\theta}_0(1) + \alpha_t \boldsymbol{\phi}_1(a) \delta_t = 5$$
$$\boldsymbol{\theta}_1(2) \leftarrow \boldsymbol{\theta}_0(2) + \alpha_t \boldsymbol{\phi}_2(a) \delta_t = 5,$$

leading to the updated parameter vector $\boldsymbol{\theta}_1 = [5, 5]^\top$. The resulting Q-function is $\hat{\mathsf{Q}}(\boldsymbol{\theta}_1) = [10, 15]$. Notice that, as expected, $\|\mathsf{Q}^\pi - \hat{\mathsf{Q}}(\boldsymbol{\theta}_1)\| < \|\mathsf{Q}^\pi - \hat{\mathsf{Q}}(\boldsymbol{\theta}_0)\|$. However, if this estimate is used in a greedy policy update, it will cause the policy to increase the probability of action b and decrease that of action a, unlike what is intended. ◇

In the example above, since the target function can be represented exactly in $\mathcal{Q}$, one would expect the algorithm to eventually settle in the correct values for $\boldsymbol{\theta}$, leading to a correct greedy policy update. However, in the general case where only an approximation is computed, the same need not happen. This is due to the fact that eligibility vectors cannot generally distinguish between the contribution of different actions to the error (given the current policy). To overcome this difficulty, we introduce the concept of *differential eligibility vector*.[4] A differential eligibility vector updates the parameter vector proportionally to the differential $\boldsymbol{\psi}(x, a) = \boldsymbol{\phi}(x, a) - \mathbb{E}_{A \sim \pi(x)}[\boldsymbol{\phi}(x, A)]$. By removing the "common component" $\mathbb{E}_{A \sim \pi(x)}[\boldsymbol{\phi}(x, A)]$, the differential $\boldsymbol{\psi}(x, a)$ is able to distinguish more accurately the contribution of different actions in each component of $\boldsymbol{\theta}$. Using the differential eligibility vectors, the TD-$Q(0)$ update rule is

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \alpha_t \delta_t \boldsymbol{\psi}(x_t, a_t). \tag{7}$$

We now return to our previous example.

Example 1 (cont.) Let us now apply TD-Q to the 1-state, 2-action example above, only now using the DEV introduced above. In this case we get $\boldsymbol{\theta}_1 = [-2.5, 0]^\top$, corresponding to the function $\hat{\mathsf{Q}}(\boldsymbol{\theta}_1) = [-2.5, -5]$. Interestingly, we now have $\|\mathsf{Q}^\pi - \hat{\mathsf{Q}}(\boldsymbol{\theta}_1)\| > \|\mathsf{Q}^\pi - \hat{\mathsf{Q}}(\boldsymbol{\theta}_0)\|$, but $\hat{\mathsf{Q}}(\boldsymbol{\theta}_1)$ can safely be used in a greedy policy update. ◇

The above example may be somewhat misleading in its simplicity, but still illustrates the idea behind the proposed modified eligibility vectors. Generalizing the above updates for $\lambda > 0$, we get the final update rule for our algorithm:

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \alpha_t \delta_t \mathbf{z}_t \tag{8}$$
$$\mathbf{z}_{t+1} \leftarrow \gamma\lambda \mathbf{z}_t + \boldsymbol{\phi}(x_{t+1}, a_{t+1}) - \boldsymbol{\varphi}(x_{t+1}), \tag{9}$$

where $\boldsymbol{\varphi}(x) = \mathbb{E}_{A \sim \pi(x)}[\boldsymbol{\phi}(x, A)]$. It is worth mentioning that the use of differential eligibility vectors implies that $\boldsymbol{\varphi}$ must be computed, which in turn requires the computation of an expectation. However, noting that $\mathcal{A}$ is assumed finite,

$$\boldsymbol{\varphi}(x) = \mathbb{E}_{A \sim \pi(x)}[\boldsymbol{\phi}(x, A)] \triangleq \sum_{a \in \mathcal{A}} \pi(x, a) \boldsymbol{\phi}(x, a),$$

which is a simple dot product between $\pi(x, \cdot)$ and $\boldsymbol{\phi}(x, \cdot)$.

Convergence of TD-$Q(\lambda)$ with DEV

We now analyze the convergence of the update (8) when a fixed policy π is used. Let $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathsf{P}, r, \gamma)$ be an MDP with compact state-space $\mathcal{X} \subset \mathbb{R}^p$ and π a given stationary policy. We assume that the Markov chain $(\mathcal{X}, \mathsf{P}_\pi)$ is geometrically ergodic with invariant measure μ and denote by μ_π the probability measure induced on $\mathcal{X} \times \mathcal{A}$ by μ and π. We assume that the basis functions $\phi_i, i = 1, \ldots, M$, are bounded and let $\mathcal{Q}$ denote its linear span. Much like TD(λ), standard TD-Q can be interpreted as a sample-based implementation of the recursion $\hat{\mathsf{Q}}(\boldsymbol{\theta}_{k+1}) = \Pi_{\mathcal{Q}} \mathbf{H}^{(\lambda)} \hat{\mathsf{Q}}(\boldsymbol{\theta}_k)$, where $\mathbf{H}^{(\lambda)}$ is the TD-Q operator,

$$\begin{aligned}(\mathbf{H}^{(\lambda)} g)(x, a) \\ = \mathbb{E}_{A_t \sim \pi(X_t)} \Bigg[\sum_{t=0}^{\infty} (\lambda\gamma)^t \big[r(X_t, A_t) + \gamma g(X_{t+1}, A_{t+1}) \\ - g(X_t, A_t) \big] \mid X_0 = x, A_0 = a \Bigg] + g(x, a),\end{aligned}$$

and $\Pi_{\mathcal{Q}}$ denotes the orthogonal projection onto $\mathcal{Q}$ w.r.t. the inner product

$$\langle f, g \rangle_\pi = \mathbb{E}_{(X,A) \sim \mu_\pi}[f(X, A) g(X, A)]. \tag{10}$$

Convergence of TD-Q follows from the fact that the composite operator $\Pi_{\mathcal{Q}} \mathbf{H}^{(\lambda)}$ is a contraction in the norm induced by the inner-product in (10) (contraction of $\mathbf{H}^{(\lambda)}$ is established in Appendix A).

To establish convergence of TD-$Q(\lambda)$ with DEV, we adopt a similar argument and closely replicate the proof in (Tsitsiklis and Van Roy 1996). As before, we let

$$\psi_i(x, a) = \phi_i(x, a) - \mathbb{E}_{A \sim \pi(x)}[\phi_i(x, A)], i = 1, \ldots, M, \tag{11}$$

[4] The designation "differential" arises from the similar concept in differential drives, where the motion of a vehicle can be decomposed into a translation component, due to the *common velocity* of both powered wheels, and a rotation component, due to the *differential velocity* between the two powered wheels.

and denote by $\boldsymbol{\Gamma}_\pi$ and $\boldsymbol{\Sigma}_\pi$ the matrices

$$\boldsymbol{\Gamma}_\pi = \mathbb{E}_{(X,A)\sim\mu_\pi}\left[\boldsymbol{\psi}(X,A)\boldsymbol{\phi}^\top(X,A)\right]$$
$$\boldsymbol{\Sigma}_\pi = \mathbb{E}_{(X,A)\sim\mu_\pi}\left[\boldsymbol{\phi}(X,A)\boldsymbol{\phi}^\top(X,A)\right].$$

Let $\bar{\gamma} = (1-\lambda)\gamma/(1-\lambda\gamma)$. We have the following result.

Theorem 1. *Let $\mathcal{M}$, π and $\mathcal{Q}$ be as defined above and suppose that $\boldsymbol{\Gamma}_\pi > \bar{\gamma}^2\boldsymbol{\Sigma}_\pi$. If the step-size sequence, $\{\alpha_t\}$, verifies the standard stochastic approximation conditions $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$, then the TD-$Q(\lambda)$ algorithm with differential eligibility vectors defined in* (8) *converges with probability 1 (w.p.1) to the parameter vector $\boldsymbol{\theta}^*$ verifying the recursive relation*

$$\boldsymbol{\theta}^* = \boldsymbol{\Gamma}_\pi^{-1}\langle\boldsymbol{\psi}, \mathbf{H}^{(\lambda)}\hat{\mathsf{Q}}(\boldsymbol{\theta}^*)\rangle_\pi. \tag{12}$$

Proof. In order to minimize the disruption of the text, we provide only a brief outline of the proof and refer to appendix A for details. The proof rests on an ordinary differential equation (o.d.e.) argument, in which the trajectories of the algorithm are shown to closely follow the o.d.e.

$$\dot{\boldsymbol{\theta}}_t = \langle\boldsymbol{\psi}, \mathbf{H}^{(\lambda)}\hat{\mathsf{Q}}(\boldsymbol{\theta}_t) - \hat{\mathsf{Q}}(\boldsymbol{\theta}_t)\rangle_\pi.$$

A standard Lyapunov argument ensures that the above o.d.e. has a globally asymptotically stable equilibrium point, thus leading to the final result. □

We conclude with two observations. First of all, $\boldsymbol{\Gamma}_\pi > \bar{\gamma}^2\boldsymbol{\Sigma}_\pi$ can always be ensured by proper choice of λ. In particular, this always holds for $\lambda = 1$. Secondly, (12) states that $\boldsymbol{\theta}^*$ is such that $\boldsymbol{\psi}^\top(x,a)\boldsymbol{\theta}^*$ approximates $\mathsf{A}^\pi(x,a)$ in the linear space spanned by the functions $\psi_i(x,a), i = 1,\ldots,M$ defined in (11). In other words, DEV lead to an approximation $\hat{\mathsf{Q}}(x,a,\boldsymbol{\theta}^*)$ of $\mathsf{Q}^\pi(x,a)$ in $\mathcal{Q}$ such that $\hat{\mathsf{Q}}(x,a,\boldsymbol{\theta}^*) - \sum_b \pi(x,b)\hat{\mathsf{Q}}(x,b,\boldsymbol{\theta}^*)$ is a good approximation of A^π in the linear span of the set $\{\psi_i, i = 1,\ldots,M\}$. This is convenient for optimization purposes.

Proposition 2. *Let $\mathcal{G}$ denote the linear span of $\{\psi_i, i = 1\ldots,M\}$, where each $\psi_i, i = 1,\ldots,M$ is as defined above, and let $\Pi_{\mathcal{G}}$ denote the orthogonal projection onto $\mathcal{G}$ w.r.t. inner product in* (10). *Then, if $\boldsymbol{\theta}^*$ is defined as in* (12) *with $\lambda = 1$,*

$$\boldsymbol{\psi}^\top(x,a)\boldsymbol{\theta}^* = \Pi_{\mathcal{G}}\mathsf{A}^\pi(x,a). \tag{13}$$

Proof. See Appendix A. □

It has been argued that policy update steps can be implemented more reliably by using the advantage function instead of the Q-function (Kakade and Langford 2002). A similar result was established in the context of policy gradient methods (Bhatnagar et al. 2007), where the minimum variance in the gradient estimate is obtained by using the advantage function instead of the Q-function.

4 Applications of TD with DEV

We now describe how DEV can be integrated in LSTD(λ) (Bradtke and Barto 1996; Boyan 2002). We also discuss the advantages of using DEV in the critic of a natural actor-critic architecture (Peters, Vijayakumar, and Schaal 2005).

Least-Squares TD(λ) with DEV

The least-squares TD(λ) algorithm (Bradtke and Barto 1996; Boyan 2002) is a "batch" version of TD(λ). The literature on LSTD is extensive and we refer to (Bertsekas 2010) and references therein for a more detailed account on this methods and variations thereof. For our purposes, we consider the trivial modification of LSTD(λ) that computes the Q-values associated with a given policy, known as LSTDQ (Lagoudakis and Parr 2003). This algorithm can easily be derived from TD-$Q(\lambda)$, again resorting to the o.d.e. method of analysis. We present a simple derivation for the case where $\lambda = 0$. Noting that the TD-$Q(0)$ algorithm closely follows the o.d.e.

$$\begin{aligned}\dot{\boldsymbol{\theta}}_t &= \langle\boldsymbol{\phi}, \mathbf{H}^{(0)}\hat{\mathsf{Q}}_{\boldsymbol{\theta}_t} - \hat{\mathsf{Q}}_{\boldsymbol{\theta}_t}\rangle_\pi \\ &= \langle\boldsymbol{\phi}, r\rangle_\pi + \langle\boldsymbol{\phi}, \gamma\mathsf{P}_\pi\boldsymbol{\phi}^\top - \boldsymbol{\phi}^\top\rangle_\pi\boldsymbol{\theta}.\end{aligned}$$

Letting $\mathbf{b} = \langle\boldsymbol{\phi}, r\rangle_\pi$ and $\mathbf{M} = \langle\boldsymbol{\phi}, \boldsymbol{\phi}^\top - \gamma\mathsf{P}_\pi\boldsymbol{\phi}^\top\rangle_\pi$, it follows that TD-$Q(0)$, upon convergence, provides the solution to the linear system $\mathbf{M}\boldsymbol{\theta} = \mathbf{b}$. LSTDQ computes a similar solution by building explicit estimates $\hat{\mathbf{M}}$ and $\hat{\mathbf{b}}$ for $\mathbf{M}$ and $\mathbf{b}$ and solving the aforementioned linear system, either directly as $\boldsymbol{\theta}^* = \hat{\mathbf{M}}^{-1}\hat{\mathbf{b}}$ or iteratively as

$$\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k - \beta(\hat{\mathbf{M}}\boldsymbol{\theta}_k - \hat{\mathbf{b}}),$$

with a suitable stepsize β (Bertsekas 2010). The above algorithm can easily be modified to accomodate DEV by noting that the structure of TD-$Q(\lambda)$ with DEV is similar to that of TD-$Q(\lambda)$. In fact, by setting $\mathbf{b}_{\mathrm{DEV}} = \langle\boldsymbol{\psi}, r\rangle_\pi$ and $\mathbf{M}_{\mathrm{DEV}} = \langle\boldsymbol{\psi}, \boldsymbol{\phi}^\top - \gamma\mathsf{P}_\pi\boldsymbol{\phi}^\top\rangle_\pi$, we again have that TD-$Q(\lambda)$ with DEV computes the solution to the linear system $\mathbf{M}_{\mathrm{DEV}}\boldsymbol{\theta} = \mathbf{b}_{\mathrm{DEV}}$.

For general λ, given an infinite sampled trajectory $\{x_t\}$ of the chain $(\mathcal{X}, \mathsf{P}_\pi)$ and the corresponding action sequence, $\{a_t\}$, the estimates $\hat{\mathbf{M}}$ and $\hat{\mathbf{b}}$ for $\mathbf{M}_{\mathrm{DEV}}$ and $\mathbf{b}_{\mathrm{DEV}}$ can be built iteratively as

$$\begin{aligned}\hat{\mathbf{M}}_{t+1} &\leftarrow \hat{\mathbf{M}}_t + \mathbf{z}_t\big(\boldsymbol{\phi}(x_t) - \gamma\boldsymbol{\phi}(x_{t+1})\big)^\top \\ \hat{\mathbf{b}}_{t+1} &\leftarrow \hat{\mathbf{b}}_t + \mathbf{z}_t r(x_t, a_t) \\ \mathbf{z}_{t+1} &\leftarrow \gamma\lambda\mathbf{z}_t + \boldsymbol{\psi}(x_t, a_t).\end{aligned}$$

The target vector $\boldsymbol{\theta}^*$ can then again be computed by solving the linear system $\mathbf{M}_{\mathrm{DEV}}\boldsymbol{\theta} = \mathbf{b}_{\mathrm{DEV}}$.

Natural Actor-Critic with DEV

In this section we describe the application of DEV in a natural actor-critic setting. We start by noting the similarity between (3) and (13), it follows that TD-$Q(\lambda)$ with DEV (or its batch version described in Section 4) is naturally suited to compute the projection of A^π onto a suitable linear space $\mathcal{G}$. In order for this result to be used in a natural actor-critic setting, it remains to show whether the functions $\psi_i, i = 1,\ldots,M$ are *compatible* in the sense of (2).

To see this, consider the parameterized family of policies

$$\pi_{\boldsymbol{\omega}}(x,a) = \frac{e^{\boldsymbol{\phi}^\top(x,a)\boldsymbol{\omega}}}{\sum_b e^{\boldsymbol{\phi}^\top(x,b)\boldsymbol{\omega}}}.$$

This is nothing more than the softmax policy associated with the function $\mathsf{Q}(\omega) \in \mathcal{Q}$. Given the above softmax policy representation, it is straightforward to note that

$$\nabla_{\boldsymbol{\omega}} \log \pi_{\boldsymbol{\omega}}(x, a) = \boldsymbol{\phi}(x, a) - \sum_b \pi_{\boldsymbol{\omega}}(x, b)\boldsymbol{\phi}(x, b) = \boldsymbol{\psi}(x, a).$$

This means that we can use the estimate from TD-$Q(\lambda)$ with DEV in a gradient update as

$$\boldsymbol{\omega}_{k+1} \leftarrow \boldsymbol{\omega}_k + \beta_k \boldsymbol{\theta}_k^*,$$

where $\{\beta_k\}$ is some positive step-size sequence and $\boldsymbol{\theta}_k^*$ is the limit in (12) obtained with the policy $\pi_{\boldsymbol{\omega}_k}$. In case of convergence, this update will find a softmax policy whose associated Q-function is constant (no further improvement is possible).

Before concluding this section, we mention that the natural-actor critic algorithm thus obtained is a variation of those described in (Bhatnagar et al. 2007; Peters, Vijayakumar, and Schaal 2005). The main difference lies in the critic used as it provides a different estimate for A^π. In particular, by using TD-Q with DEV, we do not require computing a separate value function and, by setting $\lambda = 1$, we are able to recover *unbiased natural gradient estimates*, unlike the aforementioned approaches (Bhatnagar et al. 2007).

5 Discussion

In this paper, we proposed a modification of TD-$Q(\lambda)$ specifically tailored to control settings. We introduced differential eligibility vectors (DEV), designed to allow TD-$Q(\lambda)$ to more accurately learn the relative value of the actions in an MDP, rather than their absolute value. We studied the convergence properties of the algorithm thus obtained and also discussed how DEV can be integrated within LSTDQ and natural actor-critic. Our results show that TD-$Q(\lambda)$ with DEV is able to directly approximate the advantage function without requiring estimating an auxiliary value function, allowing for immediate integration in a natural actor critic architecture. In particular, by setting $\lambda = 1$, we are able to recover unbiased estimates of the natural gradient.

From a broader point-of-view, the analysis in this paper provides an interesting perspective on reinforcement learning with function approximation. Our results can be seen as a complement to the policy gradient theorem (Sutton et al. 2000): while the latter bridges policy-gradient methods and value-based methods by establishing the dependence of the gradient on the value-function, our results establish a bridge in the complementary direction, by showing that a sensible policy update when using TD-$Q(\lambda)$ with DEV in a control setting is nothing more than a policy-gradient update.

Our approach is also complementary to other works that studied eligibility vectors in control settings, mainly in off-policy RL (Precup, Sutton, and Singh 2000; Maei and Sutton 2010). Writing the eligibility update in the general form

$$\mathbf{z}_{t+1} = \rho \mathbf{z}_t + f_t,$$

where f_t is some vector that depends on the state and action at time t, the aforementioned works manipulate ρ to compensate for the off-policy learning. In this paper, we manipulate f_t, removing the common component if the features across actions. Each of the two previous modification is aimed at a different goal, and should inclusively be possible to combine both to yield yet another eligibility update.

There are several open issues still worth exploring. First of all, although we have not discussed this issue in this paper, we expect that a fully incremental version of natural actor-critic using TD-$Q(\lambda)$ with DEV as a critic should be possible, by considering a two-time-scale implementation in the spirit of (Bhatnagar et al. 2007). Also, empirical validation of the theoretical results in this paper is still necessary.

A Proofs

Contraction Properties of $\mathbf{H}^{(\lambda)}$

Lemma 3. *Let π be a stationary policy and $(\mathcal{X}, \mathsf{P}_\pi)$ the induced chain, with invariant probability measure μ_π. Then, the operator $\mathbf{H}^{(\lambda)}$ is a contraction in the norm induced by the inner product in* (10).

Proof. The proof follows that of Lemma 4 in (Tsitsiklis and Van Roy 1996). To simplify the notation, we define the operator P_π as

$$(\mathsf{P}_\pi f)(x, a) = \mathbb{E}_{Y \sim \mathsf{P}_a(x)}\left[\sum_b \pi(Y, b) f(Y, b)\right],$$

where f is some measurable function. For $\lambda = 1$ the result follows trivially from the definition of $\mathbf{H}^{(1)}$. For $\lambda < 1$, we write $\mathbf{H}^{(\lambda)}$ in the equivalent form

$$(\mathbf{H}^{(\lambda)} g)(x, a) = (1 - \lambda) \sum_{T=0}^{\infty} \lambda^T \mathbb{E}_{A_t \sim \pi(X_t)}\left[\sum_{t=0}^{T} \gamma^t r(X_t, A_t) + \gamma^{T+1} g(X_{T+1}, A_{T+1}) \mid X_0 = x, A_0 = a\right].$$

It follows that

$$\begin{aligned}(\mathbf{H}^{(\lambda)} g_1)(x, a) &- (\mathbf{H}^{(\lambda)} g_2)(x, a) \\ &= (1 - \lambda) \sum_{t=0}^{\infty} \lambda^t \gamma^{t+1} \big(\mathsf{P}_\pi^{t+1} g_1 - \mathsf{P}_\pi^{t+1} g_2\big)(x, a).\end{aligned}$$

Noticing that $\|\mathsf{P}_\pi f\|_\pi \leq \|f\|_\pi$, where $\|\cdot\|_\pi$ denotes the norm induced by the inner-product $\langle \cdot, \cdot \rangle_\pi$, we have

$$\begin{aligned}\|\mathbf{H}^{(\lambda)} g_1 &- \mathbf{H}^{(\lambda)} g_2\|_\pi \\ &= \|(1 - \lambda) \sum_{t=0}^{\infty} \lambda^t \gamma^{t+1} \left[\mathsf{P}_\pi^{t+1} g_1 - \mathsf{P}_\pi^{t+1} g_2\right]\|_\pi \\ &\leq (1 - \lambda) \sum_{t=0}^{\infty} \lambda^t \gamma^{t+1} \|g_1 - g_2\|_\pi \\ &= \frac{(1 - \lambda)\gamma}{1 - \lambda\gamma} \|g_1 - g_2\|_\pi,\end{aligned}$$

and the conclusion follows by noting that $(1 - \lambda)\gamma < 1 - \lambda\gamma$. □

Convergence of TD-$Q(\lambda)$ with DEV (Theorem 1)

The proof essentially follows that of Theorem 2.1 in (Tsitsiklis and Van Roy 1996). In particular, the assumption of geometric ergodicity of the induced chain and the fact that the basis functions are linearly independent and square integrable w.r.t. the invariant measure for the chain ensure that the analysis of the algorithm can be established by means of an o.d.e. argument (Tsitsiklis and Van Roy 1996; Benveniste, Métivier, and Priouret 1990). The associated o.d.e. can be obtained by constructing a stationary Markov chain $\{(X_t, A_t, \mathbf{Z}_t, X_{t+1})\}$, in which (X_t, A_t, X_{t+1}) are distributed according to the induced invariant measure and $\mathbf{Z}_t$ is defined as

$$\mathbf{Z}_t = \sum_{\tau=-\infty}^{t} (\lambda\gamma)^{t-\tau} \boldsymbol{\psi}(X_\tau, A_\tau).$$

The o.d.e. then becomes

$$\begin{aligned}\dot{\boldsymbol{\theta}}_t = \mathbb{E}\Bigg[\sum_{\tau=-\infty}^{t} (\lambda\gamma)^{t-\tau} \boldsymbol{\psi}(X_\tau, A_\tau)\Big(r(X_t, A_t) \\ + \gamma \sum_{b\in\mathcal{A}} \pi(X_{t+1}, b)\hat{\mathsf{Q}}(X_{t+1}, b, \boldsymbol{\theta}_t) - \hat{\mathsf{Q}}(X_t, A_t, \boldsymbol{\theta}_t)\Big)\Bigg],\end{aligned} \tag{14}$$

where we omitted that (X_t, A_t) is distributed according to μ_π to avoid excessive cluttering the notation. By adjusting the index in the summation, the above can be rewritten as

$$\begin{aligned}h(\boldsymbol{\theta}_t) &= \mathbb{E}\Bigg[\sum_{t=0}^{\infty}(\lambda\gamma)^t \boldsymbol{\psi}(X_0, A_0)\Big[r(X_t, A_t) \\ &\qquad + \gamma \sum_{b\in\mathcal{A}} \pi(X_{t+1}, b)\hat{\mathsf{Q}}(X_{t+1}, b, \boldsymbol{\theta}_t) - \hat{\mathsf{Q}}(X_t, A_t, \boldsymbol{\theta}_t)\Big]\Bigg] \\ &= \langle \boldsymbol{\psi}, \mathbf{H}^{(\lambda)}\hat{\mathsf{Q}}(\boldsymbol{\theta}_t) - \hat{\mathsf{Q}}(\boldsymbol{\theta}_t)\rangle_\pi.\end{aligned}$$

We now establish global asymptotic stability of the o.d.e. (14). Let $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ be two trajectories of the o.d.e. (we omit the time-dependency to avoid excessively cluttering the notation). Let $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$. Then,

$$\begin{aligned}\frac{d}{dt}\|\tilde{\boldsymbol{\theta}}\|_2^2 &= 2\tilde{\boldsymbol{\theta}}^\top \big(h(\boldsymbol{\theta}_1) - h(\boldsymbol{\theta}_2)\big) \\ &= 2\langle \boldsymbol{\psi}^\top\tilde{\boldsymbol{\theta}}, \mathbf{H}^{(\lambda)}\hat{\mathsf{Q}}(\boldsymbol{\theta}_1) - \mathbf{H}^{(\lambda)}\hat{\mathsf{Q}}(\boldsymbol{\theta}_2)\rangle_\pi - 2\langle \boldsymbol{\psi}^\top\tilde{\boldsymbol{\theta}}, \boldsymbol{\phi}^\top\tilde{\boldsymbol{\theta}}\rangle_\pi\end{aligned}$$

Applying Hölder's inequality we get

$$\frac{d}{dt}\|\tilde{\boldsymbol{\theta}}\|_2^2 \leq -2\tilde{\boldsymbol{\theta}}^\top\boldsymbol{\Gamma}_\pi\tilde{\boldsymbol{\theta}} + \frac{2(1-\lambda)\gamma}{1-\lambda\gamma}\sqrt{(\tilde{\boldsymbol{\theta}}^\top\boldsymbol{\Gamma}_\pi\tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}^\top\boldsymbol{\Sigma}_\pi\tilde{\boldsymbol{\theta}})}$$

where we have used the facts that $\mathbf{H}^{(\lambda)}$ is a contraction and $\mathbb{E}_{(X,A)\sim\mu_\pi}\left[\boldsymbol{\psi}(X,A)\boldsymbol{\psi}^\top(X,A)\right] = \boldsymbol{\Gamma}_\pi$. Since, by assumption, $\boldsymbol{\Gamma}_\pi > \bar{\gamma}^2\boldsymbol{\Sigma}_\pi$, it holds that $\frac{d}{dt}\|\tilde{\boldsymbol{\theta}}\|_2^2 < 0$ and global asymptotic stability of the o.d.e. (14) follows from a standard Lyapunov argument. Convergence of w.p.1 TD-$Q(\lambda)$ with DEV follows. Finally, explicitly computing the equilibrium point of (14) leads to

$$\langle \boldsymbol{\psi}, \boldsymbol{\phi}^\top\boldsymbol{\theta}^*\rangle_\pi = \langle \boldsymbol{\psi}, \mathbf{H}^{(\lambda)}\boldsymbol{\phi}^\top\boldsymbol{\theta}^*\rangle_\pi$$

which, solving for $\boldsymbol{\theta}^*$, yields $\boldsymbol{\theta}^* = \boldsymbol{\Gamma}_\pi^{-1}\langle \boldsymbol{\psi}, \mathbf{H}^{(\lambda)}\boldsymbol{\phi}^\top\boldsymbol{\theta}^*\rangle_\pi$. □

Limit Point of TD-$Q(1)$ with DEV (Proposition 2)

The result follows from observing that, given any two functions $f, g : \mathcal{X}\times\mathcal{A} \to \mathbb{R}$,

$$\langle f - \pi f, g - \pi g\rangle_\pi = \langle f, g - \pi g\rangle_\pi = \langle f - \pi f, g\rangle_\pi, \tag{15}$$

where wrote πf to denote the function $(\pi f)(x) = \mathbb{E}_{A\sim\pi(x)}\left[f(x, A)\right]$. Using the result from Theorem 1, we have, for $\lambda = 1$

$$\boldsymbol{\theta}^* = \boldsymbol{\Gamma}_\pi^{-1}\langle \boldsymbol{\psi}, \mathbf{H}^{(1)}\boldsymbol{\phi}^\top\boldsymbol{\theta}^*\rangle_\pi = \boldsymbol{\Gamma}_\pi^{-1}\langle \boldsymbol{\psi}, \mathsf{Q}^\pi\rangle_\pi.$$

Using (15), we have

$$\boldsymbol{\theta}^* = \boldsymbol{\Gamma}_\pi^{-1}\langle \boldsymbol{\psi}, \mathsf{Q}^\pi - \pi(\mathsf{Q}^\pi)\rangle_\pi = \boldsymbol{\Gamma}_\pi^{-1}\langle \boldsymbol{\psi}, \mathsf{Q}^\pi - \mathsf{V}^\pi\rangle_\pi$$

and the result follows from the observation that $\boldsymbol{\Gamma}_\pi = \mathbb{E}_{(X,A)\sim\mu_\pi}\left[\boldsymbol{\psi}(X,A)\boldsymbol{\psi}^\top(X,A)\right]$. □

Acknowledgements

This work was supported by the Portuguese *Fundação para a Ciência e a Tecnologia* (INESC-ID multiannual funding) through the PIDDAC Program funds.

References

Antos, A.; Szepesvári, C.; and Munos, R. 2008. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Mach. Learn.* 71:89–129.

Baird, L. 1993. Advantage updating. Technical Report WL-TR-93-1146, Wright Laboratory, Wright-Patterson Air Force Base.

Baxter, J., and Bartlett, P. 2001. Infinite-horizon policy-gradient estimation. *J. Artificial Intelligence Research* 15:319–350.

Benveniste, A.; Métivier, M.; and Priouret, P. 1990. *Adaptive Algorithms and Stochastic Approximations*, vol. 22. Springer-Verlag.

Bertsekas, D. 2010. Approximate policy iteration: A survey and some new methods. Technical Report LIDS-2833, MIT.

Bhatnagar, S.; Sutton, R.; Ghavamzadeh, M.; and Lee, M. 2007. Incremental natural actor-critic algorithms. In *Adv. Neural Information Proc. Systems*, volume 20.

Boyan, J. 2002. Technical update: Least-squares temporal difference learning. *Machine Learning* 49:233–246.

Bradtke, S., and Barto, A. 1996. Linear least-squares algorithms for temporal difference learning. *Mach. Learn.* 22:33–57.

Kakade, S., and Langford, J. 2002. Approximately optimal approximate reinforcement learning. In *19th Int. Conf. Machine Learning*.

Kakade, S. 2001. A natural policy gradient. In *Adv. Neural Information Proc. Systems*, volume 14, 1531–1538.

Konda, V., and Tsitsiklis, J. 2003. On actor-critic algorithms. *SIAM J. Control and Optimization* 42(4):1143–1166.

Lagoudakis, M., and Parr, R. 2003. Least-squares policy iteration. *J. Machine Learning Research* 4:1107–1149.

Maei, H., and Sutton, R. 2010. GQ(): A general gradient algorithm for temporal-difference prediction learning with eligibility traces. In *Proc. 3rd Int. Conf. Artificial General Intelligence*, 1–6.

Marbach, P., and Tsitsiklis, J. 2001. Simulation-based optimization of Markov reward processes. *IEEE Trans. Automatic Control* 46(2):191–209.

Melo, F.; Meyn, S.; and Ribeiro, M. 2008. An analysis of reinforcement learning with function approximation. In *Proc. 25th Int. Conf. Machine Learning*, 664–671.

Meyn, S., and Tweedie, R. 1993. *Markov Chains and Stochastic Stability*. Springer-Verlag.

Perkins, T., and Precup, D. 2003. A convergent form of approximate policy iteration. In *Adv. Neural Information Proc. Systems*, vol. 15, 1595–1602.

Peters, J.; Vijayakumar, S.; and Schaal, S. 2005. Natural actor-critic. In *Proc. 16th European Conf. Machine Learning*, 280–291.

Precup, D.; Sutton, R.; and Singh, S. 2000. Eligibility traces for off-policy policy evaluation. In *Proc. 17th Int. Conf. Machine Learning*, 759–766.

Sutton, R.; McAllester, D.; Singh, S.; and Mansour, Y. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Adv. Neural Information Proc. Systems*, vol. 12.

Szepesvári, C., and Smart, W. 2004. Interpolation-based Q-learning. In *Proc. 21st Int. Conf. Machine learning*, 100–107.

Tsitsiklis, J., and Van Roy, B. 1996. An analysis of temporal-difference learning with function approximation. *IEEE Trans. Automatic Control* 42(5):674–690.