Toward Off-Policy Learning Control with Function Approximation

Hamid Reza Maei,* **Csaba Szepesvári,*** **Shalabh Bhatnagar,†** **Richard S. Sutton***
*Department of Computing Science, University of Alberta, Edmonton, Canada T6G 2E8
†Department of Computer Science and Automation, Indian Institute of Science, Bangalore-560012, India

Abstract

We present the first temporal-difference learning algorithm for off-policy control with unrestricted linear function approximation whose per-time-step complexity is linear in the number of features. Our algorithm, *Greedy-GQ*, is an extension of recent work on gradient temporal-difference learning, which has hitherto been restricted to a prediction (policy evaluation) setting, to a control setting in which the target policy is greedy with respect to a linear approximation to the optimal action-value function. A limitation of our control setting is that we require the behavior policy to be stationary. We call this setting *latent learning* because the optimal policy, though learned, is not manifest in behavior. Popular off-policy algorithms such as Q-learning are known to be unstable in this setting when used with linear function approximation.

In reinforcement learning, the term "off-policy learning" refers to learning about one way of behaving, called the *target policy*, from data generated by another way of selecting actions, called the *behavior policy*. The target policy is often an approximation to the optimal policy, which is typically deterministic, whereas the behavior policy is often stochastic, exploring all possible actions in each state as part of finding the optimal policy. Freeing the behavior policy from the target policy enables a greater variety of exploration strategies to be used. It also enables learning from training data generated by unrelated controllers, including manual human control, and from previously collected data. A third reason for interest in off-policy learning is that it permits learning about multiple target policies (e.g., optimal policies for multiple subgoals) from a single stream of data generated by a single behavior policy.

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

Off-policy learning for tabular (non-approximate) settings is well understood; there exist simple, online algorithms such as Q-learning (Watkins & Dayan, 1992) which converge to the optimal target policy under minimal conditions. For approximation settings, however, results are much weaker. One promising recent development is gradient-based temporal-difference (TD) learning methods, which have been proven stable under off-policy learning for linear (Sutton et al., 2009a) and nonlinear (Maei et al., 2010) function approximators. However, so far this work has only applied to prediction settings, in which both the target and behavior policy are stationary. In this paper we generalize prior work with gradient TD methods by allowing changes in the target policy. In particular, we consider learning an approximation to the optimal action-value function (thereby finding an approximately optimal target policy) from data generated by an arbitrary stationary behavior policy. We call this problem setting *latent learning* because the optimal policy is learned but remains latent; it is not allowed to be overtly expressed in behavior. Our latent learning result could be extended further, for example to allow the behavior policy to change slowly as long as it remained sufficiently exploratory, but it is already a significant step. Our results build on ideas from prior work with gradient TD methods but require substantially different techniques to deal with the control case.

We present a new latent learning algorithm, *Greedy-GQ*, which possesses a number of properties that we find desirable: 1) Linear function approximation; 2) No restriction on the features used; 3) Online, incremental, with memory and per-time-step computation costs that are linear in the number of features; and 4) Convergent to a local optimum or equilibrium point. Alternative ways of solving the latent learning problem include using non-incremental methods that are more computationally expensive (e.g., Lagoudakis & Parr, 2003; Antos et al., 2008; 2007), possibly with nonlinear value function approximation methods (e.g., Antos et al., 2008; 2007); putting restrictions on the linear function approximation method (Gordon, 1995; Szepesvári & Smart, 2004), or on the interaction of

the sample and the features (Melo et al., 2008). Non-incremental methods that allow non-linear value function approximation are an interesting alternative. Because they are non-incremental, there are no stability issues arising. The price is that their computational complexity is harder to control. For a discussion of the relative merits of (non-)incremental methods the reader is referred to Section 2.2.3 of (Szepesvári, 2009). Previous theoretical attempts to construct incremental methods with the above properties include that of (Szepesvári & Smart, 2004) and (Melo et al., 2008), which also discuss relevant prior literature. The first of these works suggests to use interpolative function approximation techniques (restricting the features), the second work proves convergence only in the case when the sample distribution and the features are matched in some sense. Both works prove convergence to a fixed point of a suitably defined operator.

In contrast, our algorithm is not restricted in the choice of the features. However, we are able to prove only convergence to the equilibria of a suitably defined cost function. The cost function that our algorithm attempts to minimize is the projected Bellman error (Sutton et al., 2009a) which is extended to the control setting in this paper.

1. The learning problem

We assume that the reader is familiar with basic concepts of MDPs (for a refreshment of these concepts, we refer the reader to Sutton & Barto (1998)). The purpose of this section is to define the learning problem and to define our notation.

We consider the following *latent learning* scenario: An agent interacts with its environment. The interaction results in a sequence $S_0, A_0, R_1, S_1, A_1, \dots$ of random variables, where for $t \geq 0$, $S_t \in \mathcal{S}$ are states, $A_t \in \mathcal{A}$ are actions, $R_{t+1} \in \mathbb{R}$ are rewards.[1] Fix $t \geq 0$ and let $H_t = (S_0, A_0, R_1, \dots, S_t)$ be the *history* up to time t. It is assumed that a fixed behavior policy π_b is used to generate the actions: $A_t \sim \pi_b(\cdot|S_t)$, independently of the history H_t given S_t. Thus, here for any $s \in \mathcal{S}$, $\pi_b(\cdot|s)$ is a probability distribution over $\mathcal{A}$. It is also assumed that $(S_{t+1}, R_{t+1}) \sim P(\cdot, \cdot|S_t, A_t)$, independently of H_t given S_t, A_t. Here P is the joint next-state and reward distribution kernel. For simplicity, we assume that (S_t, A_t) is in its steady-state and we use μ to denote the underlying distribution.

The goal of the agent is to learn an optimal policy for the MDP, $M = (S, A, P)$, with respect to the total expected discounted reward criterion. The optimal action-value function under this criterion shall be denoted by Q^*. As it is well known, acting greedily w.r.t. Q^* leads to an optimal policy. Remember that a policy π is *greedy* w.r.t. an action-value function Q if for every state s, π selects an action (possibly random) amongst the maximizers of $Q(s, \cdot)$. The Bellman operator acting on action-value functions underlying a stationary policy π shall be denoted by T^π, and is defined by

$$T^\pi Q(s,a) = \int \{r(s,a,s') + \gamma Q(s',b)\}\, \pi(db|s') P_{\mathcal{S}}(ds'|s,a),$$

where $r(s, a, s')$ is the expected immediate reward of transition (s, a, s'), $P_{\mathcal{S}}(\cdot|s, a)$ is the next-state distribution (the marginal of $P(\cdot, \cdot|s, a)$) and we are slightly abusing notation by using integral signs to denote both sums and integrals, depending on whether the respective spaces are discrete or continuous.

2. Derivation of Greedy-GQ

The purpose of this section is to derive the new algorithm.

We use linear value function approximation of the form $Q_\theta(s, a) = \theta^\top \varphi(s, a)$, $(s, a) \in \mathcal{S} \times \mathcal{A}$, to approximate Q^*. Here $\varphi(s, a) \in \mathbb{R}^d$ are the features, $\theta \in \mathbb{R}^d$ are the parameters to be tuned. We also employ a class of stationary policies, $(\pi_\theta; \theta \in \mathbb{R}^d)$. For each $\theta \in \mathbb{R}^d$, π_θ is a stationary policy (possibly stochastic). We will use $\pi_\theta(\cdot|s)$ to denote the action-selection probability distribution chosen by π_θ at state s. Two choices of particular interest are the greedy class and the (truncated) Gibbs class: For the *greedy class*, for any $\theta \in \mathbb{R}^d$, $\pi_\theta(\cdot|s)$ is a greedy policy w.r.t. Q_θ. For the Gibbs class, the set $\mathcal{A}$ is assumed to be countable and $\pi_\theta(a|s) \propto e^{\kappa(Q_\theta(s,a))}$, where (e.g.) $\kappa(x) = c/(1 + \exp(-x))$ with some $c > 0$.

The main idea of the algorithm is to minimize the projected Bellman error

$$J(\theta) = \|\Pi T^{\pi_\theta} Q_\theta - Q_\theta\|_\mu^2$$

using (approximate) stochastic gradient descent. Here $\|Q\|_\mu^2 = \int Q^2(s,a)\mu(da, ds)$ and Π is a projection operator which projects action-value functions into the linear space $\mathcal{F} = \{ Q_\theta : \theta \in \mathbb{R}^d \}$ w.r.t. $\|\cdot\|_\mu$: $\Pi\hat{Q} = \operatorname{argmin}_{f \in \mathcal{F}} \|\hat{Q} - f\|_\mu$.

We aim at an algorithm that works both in the case when $(\pi_\theta; \theta \in \mathbb{R}^d)$ is the greedy class, or when $(\pi_\theta; \theta \in \mathbb{R}^d)$ is a smooth class. Note that in the former case π_θ is non-differentiable w.r.t. θ, implying the lack of differentiability of J. In this case we will use subdifferentials and our method becomes an approximate stochastic subgradient method.

The motivation to minimize J is twofold: (Approximate) gradient descent lets us avoid divergence issues. When (π_θ) is the greedy class, $T^{\pi_\theta} Q_\theta = T^* Q_\theta$,

[1] To avoid measurability issues assume that $\mathcal{S}, \mathcal{A}$ are at most countably infinite. However, the results extend to more general spaces with some additional assumptions.

where T^* is the Bellman optimality operator acting on action-value functions. It can be shown that if Q-learning converged, then it would converge to the solution of $\Pi T^* Q_\theta = Q_\theta$, which defines the global optima to J. Further, J has no other global maxima. Thus, if our algorithm converged to a global maximizer of J then the limit would be the same as the limit that Q-learning would choose. We note in passing that although our objective function resembles that of (Sutton et al., 2009a) and we use some ideas of this previous work, our problem and techniques are substantially different from those of (Sutton et al., 2009a) (and similar other works), who deal with prediction problems only, while we focus on control learning.

Since we will deal with non-differentiable functions, we have to work with sub-gradients. The sub-gradient of a non-convex function is defined as:

Definition 1. (Fréchet sub-gradient): The Fréchet sub-gradient of $f : \mathbb{R}^d \to \mathbb{R}$, at $x \in \mathbb{R}^d$, denoted by $\partial f(x)$ is the set of all $u^* \in \mathbb{R}^d$ such that

$$\liminf_{h \to 0, h \neq 0} \|h\|^{-1} \left[f(x+h) - f(x) - h^\top u^* \right] \geq 0.$$

Although when the greedy policy class is used, $J(\theta)$ is not differentiable, it is still a piece-wise quadratic, continuous function which is differentiable everywhere except the boundaries between the regions defining the pieces (J is not convex, unfortunately).

In order to derive a gradient for J, we notice that we can rewrite J as [2]

$$J(\theta) = \mathbb{E}[\delta_{t+1}(\theta)\varphi_t]^\top \mathbb{E}[\varphi_t \varphi_t^\top]^{-1} \mathbb{E}[\delta_{t+1}(\theta)\varphi_t],$$

where $\varphi_t = \varphi(S_t, A_t)$ is the feature at time t,

$$\delta_{t+1}(\theta) = R_{t+1} + \gamma \overline{V}_{t+1}(\theta) - \theta^\top \varphi_t$$

is the temporal difference error, and $\overline{V}_{t+1}(\theta) = \overline{V}_\theta(S_{t+1})$ is the expected value of the next state under π_θ:

$$\overline{V}_\theta(s) = \int \theta^\top \varphi(s,a) \pi_\theta(da|s). \tag{1}$$

Due to the chain-rule of subdifferentials (e.g., Kruger, 2003), it follows that if $\hat{\varphi}_{t+1}(\theta)$ is an unbiased estimate of the subgradient of $\overline{V}_{t+1}(\theta)$ (given S_{t+1}), then $b_{t+1}(\theta) = \gamma \hat{\varphi}_{t+1}(\theta) - \varphi_t$ is a subdifferential to $\delta_{t+1}(\theta)$ and thus

$$\begin{aligned} &\mathbb{E}[\, b_{t+1}(\theta)\varphi_t^\top \,] \, \mathbb{E}\left[\varphi_t \varphi_t^\top\right]^{-1} \mathbb{E}[\, \delta_{t+1}(\theta)\varphi_t \,] = \\ &\quad = -\mathbb{E}[\, \delta_{t+1}(\theta)\varphi_t \,] + \gamma \, \mathbb{E}[\, \hat{\varphi}_{t+1}(\theta)\varphi_t^\top \,] \, w^*(\theta) \end{aligned}$$

is a subdifferential to $\frac{1}{2} J(\theta)$. Here,

$$w^*(\theta) = \mathbb{E}\left[\varphi_t \varphi_t^\top\right]^{-1} \mathbb{E}[\delta_{t+1}(\theta)\varphi_t].$$

Making use of the weight-doubling trick of Sutton et al. (2009b), we introduce a new set of weights $w_t \in \mathbb{R}^d$ to estimate $w^*(\theta_t)$. The update equations, which aim at following a negated subgradient to $J(\cdot)$, then become

$$\theta_{t+1} = \theta_t + \alpha_t \left[\delta_{t+1}(\theta_t)\varphi_t - \gamma (w_t^\top \varphi_t) \hat{\varphi}_{t+1}(\theta_t) \right], \tag{2}$$

$$w_{t+1} = w_t + \beta_t \left[\delta_{t+1}(\theta_t) - \varphi_t^\top w_t \right] \varphi_t \,, \tag{3}$$

which define our algorithm *Greedy-GQ*.

Note that if the greedy class is used, an appropriate choice for $\hat{\varphi}_{t+1}(\theta_t)$ is $\hat{\varphi}_{t+1}(\theta_t) = \varphi(S_{t+1}, A'_{t+1})$, where A'_{t+1} is some maximizing action of $Q_{\theta_t}(S_{t+1}, \cdot)$. That this holds follows from the definition of subdifferentials immediately (see, e.g., Kruger 2003).

When $\pi_\theta(a|s)$ is differentiable w.r.t. θ then $\nabla \overline{V}_\theta(s) = \int [\varphi(s,a) + Q_\theta(s,a) \nabla \ln \pi_\theta(a|s)] \, \pi_\theta(da|s)$ and $\partial \overline{V}_{t+1}(\theta) = \left\{ \nabla \overline{V}_\theta(S_{t+1}) \right\}$, i.e., the subdifferential set is a singleton. Note that when the action set is large, the algorithm can just sample $A'_{t+1} \sim \pi_{\theta_t}(\cdot|S_{t+1})$ and use $\hat{\varphi}_{t+1}(\theta_t) = \varphi(S_{t+1}, A'_{t+1}) + Q_\theta(S_{t+1}, A'_{t+1}) \psi_\theta(A'_{t+1}|S_{t+1})$, where $\psi_\theta(a|s) = \nabla \ln \pi_\theta(a|s)$ is the so-called score function underlying the policy π.

Greedy-GQ uses an update-rule for parameter θ analogous to that of Q-learning with function approximation except that we have a correction term. The update of the second set of weights, w_t, follows the least mean square (LMS) rule. These weights are normally initialized to zero. As promised, the computation of an update takes linear time in the dimension of the features, d.

The update rules of Greedy-GQ are similar to GQ(λ) with $\lambda = 0$ (Maei & Sutton, 2010). However, GQ(λ) is restricted to prediction problems, whereas the present paper considers control learning.

3. Convergence analysis

We prove our results under the following conditions. The first set of conditions concerns the data $((S_t, A_t, R_{t+1}); t \geq 0)$.

(M1) $(S_{t+1}, R_{t+1}) \sim P(\cdot, \cdot|S_t, A_t)$;

(M2) $\exists \hat{R}_{\max}$ s.t. $\mathrm{Var}\,[R_{t+1}|S_t] \leq \hat{R}_{\max}$ holds almost surely (a.s.);

(M3) $A_t \sim \pi_b(\cdot|S_t)$;

(M4) The Markov process $((S_t, A_t); t \geq 0)$ is in steady-state.[3]

[2] The derivation of this follows identical steps to the derivation of the analogous identity derived for prediction problems earlier and is thus omitted. The interested reader is referred to e.g. (Sutton et al., 2009a) for the details.

[3] Note that (M4) could be replaced by a weaker con-

We also make the following assumption on the features $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$:

(P1) $\Phi_{\max} = \|\varphi\|_\infty < +\infty$;

(P2) The matrix $C = \mathbb{E}\left[\varphi_t \varphi_t^\top\right]$ is non-singular.

For $s \in \mathcal{S}$, let $\pi_\theta(\cdot|s)$ be a probability distribution. We assume that $\hat{\varphi}_{t+1}(\theta)$ is an unbiased estimate of the subgradient of $\overline{V}_{t+1}(\theta) = \overline{V}_\theta(S_{t+1})$ (cf. (1) for the definition of $\overline{V}_\theta$):

(G1) $\mathbb{E}\left[\hat{\varphi}_{t+1}(\theta)|S_{t+1}\right] \in \partial \overline{V}_{t+1}(\theta)$.

We need the following additional assumption:

(B1) The second moment of $\hat{\varphi}_{t+1}(\theta)$ is uniformly bounded: $\sup_{\theta \in \mathbb{R}^d} \mathbb{E}\left[\|\hat{\varphi}_{t+1}(\theta)\|^2\right] < +\infty$.

Under this condition and (P1), it immediately follows that the norm of the matrix

$$B(\theta) = \mathbb{E}\left[\hat{\varphi}_{t+1}(\theta)\varphi_t^\top\right]$$

is uniformly bounded, too.

Note that when the greedy policy is used this assumption is automatically satisfied under (P1). When the policy π_θ is differentiable then it will be satisfied under (P1) provided that $\sup_{\theta \in \mathbb{R}^d, (s,a) \in \mathcal{S}\times\mathcal{A}} \|\nabla \log \pi_\theta(a|s)\| < +\infty$ also holds.

We also need the following assumption on the limiting behavior of the parametric family π_θ:

(L1) For any θ, the policy $\pi_\theta^{(\infty)}$ defined by

$$\pi_\theta^{(\infty)}(a|s) = \lim_{c\to\infty} \pi_{c\theta}(a|s), \qquad (s,a) \in \mathcal{S}\times\mathcal{A}$$

exists and the convergence is uniform on compact sets.

(L2) The set $\mathcal{L} = \left\{ \pi_\theta^{(\infty)} : \theta \in \mathbb{R}^d \right\}$ is finite.

(L3) The matrices $C - \gamma \int \varphi(s',b)\varphi(s,a)^\top \pi(db|s') P_{\mathcal{S}}(ds'|s,a)\mu(ds,da)$ are non-singular, for any $\pi \in \mathcal{L}$.

Condition (L1) is satisfied for the typical choices of policy classes. Note that if π_θ is the greedy policy then (L1) is automatically satisfied. Condition (L2) will be naturally satisfied in finite state-action MDPs.

dition on the Harris recurrence of this Markov process. The modifications to our analysis would be standard (Szepesvári & Smart (2004) used this condition in a similar context). The reason for relying on (M4) is to keep matters relatively simple.

This is a technical condition that we believe can be relaxed. It is used only in the proof of the boundedness of the iterates. Condition (L3) is similar to the feature-independence condition. If it is not satisfied, the equilibrium set of J in fact can be unbounded (which does not affect value convergence, but affects the boundedness of parameters).

Now, write the algorithm in the form

$$\theta_{t+1} = \theta_t + \alpha_t\, G_{t+1}(\theta_t, w_t), \tag{4a}$$
$$w_{t+1} = w_t + \beta_t\, H_{t+1}(\theta_t, w_t), \tag{4b}$$

where

$$\begin{aligned} G_{t+1}(\theta,w) &= \delta_{t+1}(\theta)\varphi_t - \gamma\hat{\varphi}_{t+1}(\theta)\varphi_t^\top w, \\ H_{t+1}(\theta,w) &= \delta_{t+1}(\theta)\varphi_t - \varphi_t\varphi_t^\top w, \\ \delta_{t+1}(\theta) &= R_{t+1} + \gamma\,\overline{V}_{t+1}(\theta) - \theta^\top\varphi_t, \\ \varphi_t &= \varphi(S_t, A_t). \end{aligned}$$

We will use the following assumptions on the step-size sequences:

(S1) $\alpha_t, \beta_t > 0\ \forall t$ and are deterministic;

(S2) $\sum_{t=0}^\infty \alpha_t = \sum_{t=0}^\infty \beta_t = +\infty$;

(S3) $\sum_{t=0}^\infty (\alpha_t^2 + \beta_t^2) < +\infty$;

(S4) $\alpha_t/\beta_t \to 0$.

The last assumption puts the update into the class of two timescale stochastic approximation algorithms.

Define the mean update directions $g(\theta,w) = \mathbb{E}\left[G_{t+1}(\theta,w)\right]$ and $h(\theta,w) = \mathbb{E}\left[H_{t+1}(\theta,w)\right]$ [4] and the noise sequences $V_{t+1} = G_{t+1}(\theta_t,w_t) - g(\theta_t,w_t)$, $U_{t+1} = H_{t+1}(\theta_t,w_t) - h(\theta_t,w_t)$, $t \geq 0$. With these choices the algorithm takes the form

$$\begin{aligned} \theta_{t+1} &= \theta_t + \alpha_t\left[g(\theta_t,w_t) + V_{t+1}\right], \\ w_{t+1} &= w_t + \beta_t\left[h(\theta_t,w_t) + U_{t+1}\right]. \end{aligned}$$

We need results on such stochastic approximation algorithms when the mean update direction is discontinuous because g depends on $\mathbb{E}\left[\hat{\varphi}_{t+1}(\theta)|S_t = s, A_t = a\right]$, which might be a discontinuous function of θ (for $(s,a) \in \mathcal{S}\times\mathcal{A}$ fixed). These results are listed in Appendix A. The main result of this paper is the following theorem:

Theorem 1. *Under the conditions listed in this section the iterates updated by Greedy-GQ stay bounded. Further, θ_t converges to $M_0 = \{\theta : 0 \in \partial J(\theta)\}$ with probability one.*

[4] These are well defined thanks to (M4).

The plan of the analysis of the algorithm is as follows: We make the working hypothesis that the parameters updated by the algorithm remain bounded almost surely:

$$\sup_t(\|\theta_t\| + \|w_t\|) < +\infty, \text{a.s.} \tag{6}$$

Then, under this assumption we show that the limiting behavior of the iterates can be reduced to that of an appropriately defined differential equation. Next, we study the limiting behavior of this differential equation. The analysis is finished by showing that (6) indeed holds. Note that by assuming further structure on J (i.e., when (π_θ) is the greedy class) and that the "noise" is sufficiently rich (i.e., it "excites" every direction), one can show that θ_t will converge to local minima of J.[5]

In what follows we will always assume that (M1)–(M4), (P1)–(P2), (G1), (L1)–(L3), (B1), (S1)–(S4) hold, so these conditions will be omitted from the results that follow.

3.1. Convergence to an invariant set

We have the following result:

Proposition 2. *Under (6), we have*

$$(\theta_t, w_t) \to \{\, (\theta, w^*(\theta)) : \theta \in M \,\}, \quad a.s.$$

Here the set $M = M(\omega) \subset \mathbb{R}^d$ is a possibly random set for which it holds almost surely that it is a compact, connected invariant set to the differential inclusion $\dot\theta(t) \in \partial J(\theta)$.

Proof. We apply Theorem 5, identifying the master equation with the update of θ_t and the slave equation with the update of w_t. We need to verify that the conditions (D5–1)–(D5–3), (S5–1)-(S5–3) and (A5–1) of Theorem 5 hold.

For this, note first that $h(\theta, w) = b - A(\theta)\theta - Cw$, where $b = \int r(s,a,s')\varphi(s,a)\, P_{\mathcal{S}}(ds'|s,a)\mu(ds,da)$, $A(\theta) = C - \gamma \int \varphi(s',b)\varphi(s,a)^\top \pi_\theta(db|s')P_{\mathcal{S}}(ds'|s,a)\mu(ds,da)$, and C was defined in (P2). Here μ denotes the stationary distribution underlying (S_t, A_t).

Now, let us verify if Condition (D5–2), which is a linear growth condition, holds for h. We have $\|h(\theta,w)\| \le \|b\| + \|A(\theta)\|\,\|\theta\| + \|C\|\|w\|$. Thus, the condition follows since $\|A(\theta)\| \le \sup_\theta \|A(\theta)\| < +\infty$, thanks to (P1). Further, for θ fixed, $h(\theta, w)$ is Lipschitz with Lipschitz constant $\|C\|$. Hence, it satisfies (D5–3).

With the help of b, $A(\theta)$ and $B(\theta)$, g can be written as $g(\theta, w) = b - A(\theta)\theta - \gamma B(\theta)w$. We see that the growth condition (D5–1) is satisfied thanks to Assumption (B1).

Now, (S5–1) is verified thanks to (S1)–(S4). To verify (S5–2), note that if $\mathcal{F}_t = \sigma(\theta_s, w_s; s \le t)$ then, thanks to their constructions, $\mathbb{E}[V_{t+1}|\mathcal{F}_t] = 0$, $\mathbb{E}[U_{t+1}|\mathcal{F}_t] = 0$. Finally, (S5–3) is verified as follows: $\|G_{t+1}(\theta,w)\| \le |\delta_{t+1}(\theta)|\,\|\varphi_t\| + \gamma\|\hat\varphi_{t+1}(\theta)\|\,\|w\|\,\|\varphi_t\| \le \Phi_{\max}(|\delta_{t+1}(\theta)| + \|\hat\varphi_{t+1}(\theta)\|\,\|w\|)$, where we used (P1). Thanks to the definition of $\delta_{t+1}(\theta)$, $|\delta_{t+1}(\theta)| \le |R_{t+1}| + \gamma|\overline{V}_{t+1}(\theta)| + \|\theta\|\,\|\varphi_t\|$. From the definition of $\overline{V}_\theta(s)$, we also get $\overline{V}_\theta(s) \le \Phi_{\max}\|\theta\|$. Hence, $|\delta_{t+1}(\theta)| \le |R_{t+1}| + 2\Phi_{\max}\|\theta\|$. By chaining the inequalities obtained and then using $(\sum_{j=1}^2 a_i)^2 \le 2\sum_{i=1}^2 a_i^2$, we get $\|G_{t+1}(\theta,w)\|^2 \le 2\Phi_{\max}^2(|R_{t+1}|^2 + 4\Phi_{\max}^2\|\theta\|^2 + \|\hat\varphi_{t+1}(\theta)\|^2\,\|w\|^2)$. Taking expectations and using (M2) and (B1), we get that $\mathbb{E}\left[\|G_{t+1}(\theta_t,w_t)\|^2\,|\mathcal{F}_t\right] \le K'(1 + \|\theta_t\|^2 + \|w_t\|^2)$ with a suitable constant $K' > 0$.

It remains to check (A5–1), i.e., if $\dot w(t) = h(\theta, w(t))$ admits a unique, globally asymptotically stable equilibrium, $w^*(\theta)$, given any fixed value of θ. Since C is a positive definite matrix, this is immediate. Further, $w^*(\theta) = C^{-1}(b - A(\theta)\theta)$, where C^{-1} exists thanks to (P2). It is immediate that $w^*(\cdot)$ is a Lipschitz continuous function since, as discussed before, $\sup_\theta \|A(\theta)\| < +\infty$. This finishes the verification of the conditions of Theorem 5.

Thus, we conclude that there exists a set $M = M(\omega) \subset \mathbb{R}^d$, which is (almost surely) a compact, connected invariant set to $\dot\theta(t) = g(\theta(t), w^*(\theta(t)))$ and $(\theta_t, w_t) \to \{\,(\theta, w^*(\theta)) : \theta \in M\,\}$, a.s. Since by construction $g(\theta, w^*(\theta)) \in -\frac{1}{2}\,\partial J(\theta)$ holds for any $\theta \in \mathbb{R}^d$, the statement follows. □

3.2. The study of the invariant set

Proposition 3. *Let M be a bounded invariant set to the differential inclusion $\dot\theta(t) \in -\frac{1}{2}\partial J(\theta)$. Then M is a subset of the set of stationary points $\mathcal{S} = \{\,\theta : 0 \in \partial J(\theta)\,\}$ to J.*

Proof. The statement is immediate when J is differentiable. When J is not differentiable, the solutions are defined in the sense of Filippov (1988) and a more careful analysis is needed. This is however omitted due to the lack of space. □

3.3. Boundedness

Proposition 4. *The iterates remain bounded, that is (6) holds.*

[5] Such assumptions are in fact necessary in the analysis of stochastic gradient descent when the objective function is non-convex (cf. Section 4.3 "Avoidance of traps" of (Borkar, 2008)). Note that the standard way to deal with this is to add noise to the updates, which would also work in our case.

Proof. We use Theorem 6. Since we have already verified the conditions of Theorem 5, it remains to show that the extra conditions of Theorem 6 hold.

Consider the function g. We need to show the existence of g_∞ such that

$$\lim_{c\to\infty} \frac{g(c\theta, w^*(c\theta))}{c} \in g_\infty(\theta)$$

where the convergence is uniform. We also need that $g_\infty(\theta)$ is such that zero is the unique global exponentially stable equilibrium to the differential inclusion

$$\dot\theta \in g_\infty(\theta). \tag{7}$$

We know that $f(\theta, w^*(\theta)) \in -\frac{1}{2}\partial J(\theta)$. Since $\partial J(c\theta) = c\partial J(c\theta)$, $c^{-1} f(c\theta, w^*(c\theta)) = c^{-2}\partial J(c\theta)$. Using the definition of J, we get

$$\frac{\partial J(c\theta)}{c^2} = \partial \left\| \frac{b}{c} - A(c\theta)\,\theta \right\|^2_{C^{-1}}. \tag{8}$$

Let

$$A_\infty(\theta) = \lim_{c\to\infty} A(c\theta). \tag{9}$$

Note that

$$A_\infty(\theta) = C - \gamma \int \varphi(s',b)\varphi(s,a)^\top \pi_\theta^{(\infty)}(db|s') P_{\mathcal{S}}(ds'|s,a)\mu(ds,da)$$

exists and the convergence in (9) is uniform on compact sets thanks to (L1). Now, take the limit of $c\to\infty$ in (8). Thanks to (L2), the interchange of limit and subdifferentials is justified and we have

$$\begin{aligned}\lim_{c\to\infty} \frac{\partial J(c\theta)}{c^2} &= \partial \| A_\infty(\theta)\theta \|^2_{C^{-1}} \\ &= \overline{\text{co}}\,\{ N(\theta)\theta \,:\, N(\theta) = \lim_{t'\to\theta} N_\infty(\theta') \},\end{aligned}$$

where $N_\infty(\theta) = 2A_\infty(\theta)^\top A_\infty(\theta)$. Note that A_∞ is piecewise constant, hence so is N_∞. Let $\{N_1,\dots,N_K\} = \{ N_\infty(\theta) : \theta\in\mathbb{R}^d \}$ and partition $\mathbb{R}^d$ into non-overlapping regions R_i, $i=1,\dots,K$, such that $N_i = N_\infty(\theta)$ for all $\theta\in R_i$.

Notice that the matrices N_i are all normal. Further, they are positive definite, because $N_i = M_i^\top M_i$ for some nonsingular matrix M_i, thanks to (L3). Further, by definition, the solutions to (7) are exactly the same as that of the switched linear system with dynamics

$$\dot\theta = -\sum_{i=1}^K \mathbb{I}_{\{\theta\in R_i\}} N_i\theta. \tag{10}$$

Thus, it suffices to study the latter system. By Lemma 2 of (Zhai et al., 2006), $\exists\rho>0$ s.t.

$$N_i + N_i^\top \succ 2\rho\mathbf{I}, \tag{11}$$

holds for $i=1,\dots,K$. Consider the Lyapunov function candidate $V(\theta) = \frac{1}{2}\theta^\top\theta$. Let $\theta(t)$ be a solution to (10). Take t such that $\dot\theta(t)$ exists. By the definition of Filippov solutions, there exists $\mu_i(t)\ge 0$, such that $\sum_{i:\theta(t)\in\overline{R}_i}\mu_i(t) = 1$ and $\dot\theta(t) = \sum_{i:\theta(t)\in\overline{R}_i}\mu_i(t)N_i\theta(t)$. Hence,

$$\begin{aligned}\dot V &= \frac{1}{2}\left(\theta(t)^\top\dot\theta(t) + \dot\theta(t)^\top\theta(t)\right) \\ &= -\frac{1}{2}\sum_{i:\theta(t)\in\overline{R}_i}\mu_i(t)\left\{\theta(t)^\top N_i\theta(t) + \theta(t)^\top N_i^\top\theta(t)\right\} \\ &\le -\rho\|\theta(t)\|^2,\end{aligned}$$

where the last inequality follows from (11). Hence, V is a Lyapunov function to (10). From (10) it is clear that zero is the only equilibrium point. Further, because $\dot V(t)\le -2\rho V(t)$, integrating both sides yields $\|\theta(t)\|^2 = 2V(t)\le C\exp(-2\rho t)$ for some $C>0$. Therefore, zero is the unique globally exponentially asymptotically stable equilibrium to (10) and thus also to (7).

Hence, we have verified all the conditions of Theorem 6 and it follows that the parameters stay uniformly bounded with probability one. □

4. Solving Baird's counterexample on Q-learning

In this section, we illustrate the convergence result of Greedy-GQ on a well known off-policy example; Baird's counterexample (Baird, 1995), for which Q-learning diverges. This has been demonstrated in Fig.1. Here, we have used the 7-star version of the

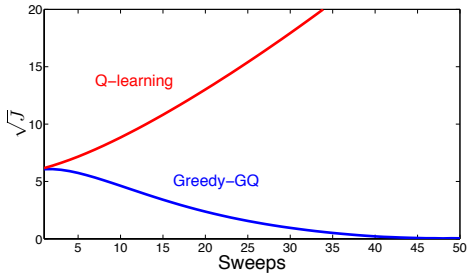

Figure 1. Empirical illustration for Baird's counterexample. The graph shows that Greedy-GQ converges to the true solution, while Q-learning diverges.

"star" counterexample. The MDP consists of 7 states and 2 actions for each state. The reward is always zero and the discount factor is $\gamma = 0.99$. In this problem, the true action value is zero for all state–action pairs. The initial value of θ parameters for the action that causes transition to the 7th state is $(1,1,1,1,1,1,10,1)$ and the rest are 1. The initial values for auxiliary weights w were set to zero. Updating

was done synchronously in dynamic-programming-like sweeps through the state-action space. The step-size parameter $\alpha = 0.1$ was used for Q-learning, and for Greedy-GQ we used $\alpha = 0.05$, $\beta = 0.25$. Fig.1 shows how the performance measure, $\sqrt{J}$, evolves with respect to the number of updates. Both algorithms used expected updates. The graph shows that Greedy-GQ finds the optimal weights, while Q-learning diverges.

Here, the choice of step-sizes goes beyond our theoretical conditions, testifying that our results are robust beyond what we can prove. For α, β converging to zero according to our theorem statement, the graphs would not differ in their behavior from the one that we presented here.

5. Conclusions and future work

In this paper we have made significant progress toward solving a long-standing open problem in reinforcement learning: the problem of off-policy learning control. Our new algorithm, Greedy-GQ, achieves the four desirable properties identified in the introduction (linear approximation, unrestricted features, an online, incremental, linear-complexity implementation, and convergence to an optimum or equilibria) in the latent-learning setting. On the other hand, our result is limited in several ways. First, we focused on the case when the behavior policy is fixed. Although this is an important case, better performance can be expected if one is allowed to actively change the way the data is sampled. Next, the algorithm might converge to local optima. This follows from the nature of the objective function considered. Unfortunately, convergence to local optima might make it difficult to derive performance bounds on the resulting policy. Nevertheless, we think that the approach considered here is a significant step towards a practical, incremental algorithm to learn a good control policy in the difficult off-policy setting. Our future plans involve extensive testing of the algorithm on various test domains and its possible extensions to prevent convergence to local minima and to handle the case when the behavior policy is allowed to change sufficiently slowly.

Acknowledgements

The authors gratefully acknowledge the insights and assistance they have received from Doina Precup, Eric Wiewiora, and David Silver. They also thank the anonymous reviewers for their helpful comments. This research was supported by iCORE and Alberta Ingenuity, both part of Alberta Innovates –Technology Futures, NSERC, MITACS and the PASCAL2 Network of Excellence under EC grant no. 216886. Cs. Szepesvári is on leave from MTA SZTAKI.

A. Results on stochastic approximation

The results here are extensions of various results in (Borkar, 2008) and can be proved using the same techniques as developed there. For brevity, the proofs of these technical results are omitted. The first result is an extension of previous two timescale stochastic approximation results where the update functions on the right-hand side (RHS) might be discontinuous.

Consider the stochastic approximation algorithm

$$x_{n+1} = x_n + a(n)\left[h(x_n, y_n) + M^{(1)}_{n+1}\right], \quad (12a)$$

$$y_{n+1} = y_n + b(n)\left[g(x_n, y_n) + M^{(2)}_{n+1}\right], \quad (12b)$$

where $x_n \in \mathbb{R}^d$, $y_n \in \mathbb{R}^k$ and $x_0 \in \mathbb{R}^d, y_0 \in \mathbb{R}^k$ are fixed (non-random), $(a(n), b(n); n \geq 0)$ are step-size sequences, $h, g : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}^d$ are possibly discontinuous, functions. As before, $(M^{(1)}_n, M^{(2)}_n; n \geq 1)$ is a noise sequence.

We shall assume that $b(n) = o(a(n))$, separating the speed at which x_n is updated from that of the update of y_n, making the algorithm a two timescale algorithm. In fact, because of this assumption the update of y_n is much smaller than the update of x_n. In the limit, we can think of that by the time y_n is updated x_n has already converged. For this reason, the update equation for y_n is called the *master update equation*, while the update for x_n is called the *slave update equation*. Analogously, y_n (x_n) is called the master (resp., slave) parameter. The above intuition suggests that if for any fixed value of $y_n = y$ the slave equation converges to some point $\lambda(y)$ fast enough then it will be sufficient to analyze the ordinary differential equation $\dot{y} = g(\lambda(y), y)$ to understand the limiting behavior of y_n. The following theorem makes this intuition precise. We shall need the following notation for this theorem: Let $f : \mathbb{R}^p \to \mathbb{R}^q$. Then, for $x \in \mathbb{R}^p$ let $\mathrm{Lim}_f(x) = \cap_{\varepsilon>0}\overline{\mathrm{co}}\left(\{f(x') \,|\, \|x - x'\| < \varepsilon\}\right)$ be the closed convex set spanned by the limit-values of f at x. Here $\overline{\mathrm{co}}(H)$ denotes the closed convex hull of set $H \subset \mathbb{R}^d$.

Theorem 5. *Consider the coupled equations (12a)–(12b). For $(x, y) \in \mathbb{R}^d \times \mathbb{R}^k$, let $G(x, y) = \mathrm{Lim}_{g(x,\cdot)}(y)$, $H(x, y) = \mathrm{Lim}_{h(\cdot,y)}(x)$. Let the following assumptions hold: $\exists K > 0$ s.t. for all $(x, y) \in \mathbb{R}^d \times \mathbb{R}^k$,*

(D5–1) $\sup_{g \in G(x,y)} \|g\| \leq K(1 + \|x\| + \|y\|)$;

(D5–2) $\sup_{h \in H(x,y)} \|h\| \leq K(1 + \|x\| + \|y\|)$;

(D5–3) h is Lipschitz in its first argument, uniformly w.r.t. the second.

Further, assume that the step-size and noise sequences satisfy:

(S5–1) $\sum_{n=0}^{\infty} a(n) = \sum_{n=0}^{\infty} b(n) = \infty$, $\sum_{n=0}^{\infty}(a(n)^2 + b(n)^2) < +\infty$, $\frac{b(n)}{a(n)} \to 0$, $n \to \infty$ *and* $(a(n))$, $(b(n))$ *are eventually decreasing;*

(S5–2) *for all* $n \geq 0$, $i = 1, 2$, $\mathbb{E}\left[M_{n+1}^{(i)} | \mathcal{F}_n\right] = 0$, *where* $\mathcal{F}_n = \sigma(x_m, y_m, M_m^{(1)}, M_m^{(2)}; m \leq n)$;

(S5–3) $\exists K' > 0$ *s.t. for all* $n \geq 0$, $i = 1, 2$, $\mathbb{E}\left[\|M_{n+1}^{(i)}\|^2 | \mathcal{F}_n\right] \leq K'(1 + \|x_n\|^2 + \|y_n\|^2)$.

In addition, assume that

(A5–1) *there exists a Lipschitz map* $\lambda : \mathbb{R}^k \to \mathbb{R}^d$ *such that for any* $y \in \mathbb{R}^k$, $\lambda(y)$ *is the globally asymptotically (uniformly) stable equilibrium to* $\dot{x}(t) = h(x(t), y)$.

Then under $\sup_n(\|x_n\| + \|y_n\|) < +\infty$, *a.s., it holds that there exists a (random, i.e., path-dependent) subset* M *of* $\mathbb{R}^d$ *such that with probability one* M *is a compact, connected and internally chain transitive invariant set to the differential equation*

$$\dot{y} = g(\lambda(y), y), \tag{13}$$

such that $(x_n, y_n) \to \hat{M} = \{\, (\lambda(y), y) \,:\, y \in M \,\}$ *a.s.*

Note that $\hat{M}$ is a random set. Also, the RHS of (13) is possibly discontinuous. When this is the case then the solutions are understood in the Filippov sense (see Filippov 1988).

The final general result concerns the boundedness of the iterates of two timescale algorithms.

Theorem 6. *Consider the update equations* (12a)–(12b). *Assume that in addition to the conditions (D5–1)–(D5–3), (S5–1)-(S5–3), (A5–1) of Theorem 5,*

1. $\exists G_\infty : \mathbb{R}^k \to 2^{\mathbb{R}^k}$ *s.t. for any* $y \in \mathbb{R}^k$, $G_\infty(y)$ *is closed convex and* $\lim_{c\to\infty} \inf_{y' \in G_\infty(y)} \|y' - \frac{g(\lambda(cy), cy)}{c}\| = 0$ *and the convergence is uniform on compacta;*

2. *Zero is the unique, globally asymptotically exponentially stable equilibrium to*

$$\dot{y} \in G_\infty(y).$$

Then, $\sup_n(\|x_n\| + \|y_n\|) < +\infty$ *holds almost surely.*

References

Antos, A., Munos, R., and Szepesvári, Cs. Fitted Q-iteration in continuous action-space MDPs. In *NIPS-20*, pp. 9–16. MIT Press, 2007.

Antos, A., Szepesvári, Cs., and Munos, R. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, April 2008.

Baird, L. C. Residual algorithms: Reinforcement learning with function approximation. In Prieditis, A. and Russell, S.J. (eds.), *ICML 1995*, pp. 30–37. IMLS, Morgan Kaufmann, 1995.

Borkar, V. S. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

Filippov, A.F. *Differential equations with discontinuous righthand sides*. Kluwer Academic Press, 1988.

Gordon, G. J. Stable function approximation in dynamic programming. In Prieditis, A. and Russell, S.J. (eds.), *ICML 1995*, pp. 261–268. IMLS, Morgan Kaufmann, 1995.

Kruger, A.Ya. On Fréchet subdifferentials. *J. of Math. Sciences*, 116:3325–3558, 2003.

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

Maei, H. R. and Sutton, R. S. GQ(λ): A general gradient algorithm for temporal-difference prediction learning with eligibility traces. In Baum, E., Hutter, M., and Kitzelmann, E. (eds.), *AGI 2010*, pp. 91–96. Atlantis Press, 2010.

Maei, H.R., Szepesvári, Cs., Bhatnagar, S., Silver, D., Precup, D., and Sutton, R.S. Convergent temporal-difference learning with arbitrary smooth function approximation. In *NIPS-22*, pp. 1204–1212, 2010.

Melo, F. S., Meyn, S. P., and Ribeiro, M. I. An analysis of reinforcement learning with function approximation. In Cohen, W. W., McCallum, A., and Roweis, S. T. (eds.), *ICML 2008*, pp. 664–671. ACM, 2008.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. Bradford Book. MIT Press, 1998.

Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, Cs., and Wiewiora, E. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In Bottou, L. and Littman, M. (eds.), *ICML 2009*, pp. 993—1000. ACM, 2009a.

Sutton, R. S., Szepesvári, Cs., and Maei, H. R. A convergent $O(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L. (eds.), *NIPS-21*, pp. 1609–1616. MIT Press, 2009b.

Szepesvári, Cs. Reinforcement learning algorithms for MDPs – a survey. Technical Report TR09-13, Department of Computing Science, University of Alberta, 2009.

Szepesvári, Cs. and Smart, W. D. Interpolation-based Q-learning. In Brodley, Carla E. (ed.), *ICML 2004*, pp. 791–798. ACM, 2004.

Watkins, C. J. C. H. and Dayan, P. Q-learning. *Machine Learning*, 3(8):279–292, 1992.

Zhai, G., Xu, X., Lin, H., and Michel, A. Analysis and design of switched normal systems. *Nonlinear Analysis*, 65:2248—2259, 2006.