Rates of Convergence of Performance Gradient Estimates Using Function Approximation and Bias in Reinforcement Learning

Gregory Z. Grudic  
University of Colorado, Boulder  
grudic@cs.colorado.edu

Lyle H. Ungar  
University of Pennsylvania  
ungar@cis.upenn.edu

Abstract

We address two open theoretical questions in Policy Gradient Reinforcement Learning. The first concerns the efficacy of using function approximation to represent the state action value function, $Q$. Theory is presented showing that linear function approximation representations of $Q$ can degrade the rate of convergence of performance gradient estimates by a factor of $O(ML)$ relative to when no function approximation of $Q$ is used, where $M$ is the number of possible actions and $L$ is the number of basis functions in the function approximation representation. The second concerns the use of a bias term in estimating the state action value function. Theory is presented showing that a non-zero bias term can improve the rate of convergence of performance gradient estimates by $O(1 - (1/M))$, where $M$ is the number of possible actions. Experimental evidence is presented showing that these theoretical results lead to significant improvement in the convergence properties of Policy Gradient Reinforcement Learning algorithms.

1 Introduction

Policy Gradient Reinforcement Learning (PGRL) algorithms have recently received attention because of their potential usefulness in addressing large continuous reinforcement Learning (RL) problems. However, there is still no widespread agreement on how PGRL algorithms should be implemented. In PGRL, the agent’s policy is characterized by a set of parameters which in turn implies a parameterization of the agent’s performance metric. Thus if $\theta \in \mathbb{R}^d$ represents a $d$ dimensional parameterization of the agent’s policy and $\rho$ is a performance metric the agent is meant to maximize, then the performance metric must have the form $\rho(\theta)$ [6]. PGRL algorithms work by first estimating the performance gradient (PG) $\partial \rho/\partial \theta$ and then using this gradient to update the agent’s policy using:

$$\theta_{t+1} = \theta_t + \alpha \frac{\partial \rho}{\partial \theta}$$

where $\alpha$ is a small positive step size. If the estimate of $\partial \rho/\partial \theta$ is accurate, then the agent can climb the performance gradient in the $\theta$ parameter space, toward locally optimal policies. In practice, $\partial \rho/\partial \theta$ is estimated using samples of the state action value function $Q$. The PGRL formulation is attractive because 1) the parameterization $\theta$ of the policy can directly imply
gests that the convergence properties of PGRL algorithms can be improved by using a bias term that is the average of \(\mathcal{C}_9\)

should avoid the use of linear FA techniques to represent \(\mathcal{C}_9\)

present theory showing that using linear basis function representations of convergence to locally optimal solutions \([5, 4]\). However, whether linear FA representations actually improves the convergence properties of PGRL is an open question. We suggested as a way of improving convergence properties. It has been proven that specific linear FA formulations can be incorporated into PGRL algorithms, while still guaranteeing convergence to locally optimal solutions \([5, 4]\). However, whether linear FA representations actually improves the convergence properties of PGRL is an open question. We present theory showing that using linear basis function representations of \(\mathcal{Q}\), rather than direct observations of it, can slow the rate of convergence of PG estimates by a factor of \(O(ML)\) (see \textbf{Theorem 1} in Section 3.1). This result suggests that PGRL formulations should avoid the use of linear FA techniques to represent \(\mathcal{Q}\). In Section 4, experimental evidence is presented supporting this conjecture.

The second open theoretical question addressed here is can a non-zero bias term \(b(s)\) in (2) improve the convergence properties of PG estimates? There has been speculation that an appropriate choice of \(b(s)\) can improve convergence properties \([6, 5]\), but theoretical support has been lacking. This paper presents theory showing that if \(b(s) = (1/M)\sum \mathcal{Q}(s,a)\), where \(M\) is the number of actions, then the rate of convergence of the PG estimate is improved by \(O(1 - (1/M))\) (see \textbf{Theorem 2} in Section 3.2). This suggests that the convergence properties of PGRL algorithms can be improved by using a bias term that is the average of \(\mathcal{Q}\) values in each state. Section 4 gives experimental evidence supporting this conjecture.

2 The RL Formulation and Assumptions

The RL problem is modeled as a Markov Decision Process (MDP). The agent’s state at time \(t \in \{1, 2, \ldots\}\) is given by \(s_t \in S, S \subseteq \mathbb{R}^P\). At each time step the agent chooses from a finite set of \(M > 1\) actions \(a_t \in A = a_1, \ldots, a_M\) and receives a reward \(r_t \in \mathbb{R}\). The dynamics of the environment are characterized by transition probabilities \(P^a_{s_{t+1}} = P_{s_t = s, a_t = a}^{n_{s', s}}\) and expected rewards \(R_s = E\{r_{t+1}|s_t = s, a_t = a\}, \forall s, s' \in S, a \in A\). The policy followed by the agent is characterized by a parameter vector \(\theta \in \mathbb{R}^d\), and is defined by the probability distribution \(\pi(s, a; \theta) = P_r\{a_t = a|s_t = s, \theta\}, \forall s \in S, a \in A\). We
assume that $\pi(s,a;\theta)$ is differentiable with respect to $\theta$.

We use the Policy Gradient Theorem of Sutton et al. [5] and limit our analysis to the start state discount reward formulation. Here the reward function $\rho(\pi)$ and state action value function $Q^{\pi}(s,a)$ are defined as:

$$\rho(\pi) = E \left\{ \sum_{t=1}^{\infty} \gamma^t r_t \mid s_0, \pi \right\}, \quad Q^{\pi}(s,a) = E \left\{ \sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k} \mid s_t = s, a_t = a, \pi \right\}$$

where $0 < \gamma \leq 1$. Then the exact expression for the performance gradient is:

$$\frac{\partial \rho}{\partial \theta} = \sum_s d^{\pi}(s) \sum_{i=1}^{M} \frac{\partial \pi(s,a_i;\theta)}{\partial \theta} (Q^{\pi}(s,a_i) - b(s)) \quad (3)$$

where $d^{\pi}(s) = \sum_{i=0}^{\infty} \gamma^i \Pr \{ s_t = s \mid s_0, \pi \}$ and $b(s) \in \mathbb{R}$.

This policy gradient formulation requires that the state-action value function, $Q^{\pi}$, under the current policy be estimated. This estimate, $\hat{Q}^{\pi}$, is derived using the observed value $Q^{\pi}_{\text{obs}}(s,a_i)$. We assume that $Q^{\pi}_{\text{obs}}(s,a_i)$ has the following form:

$$Q^{\pi}_{\text{obs}}(s,a_i) = Q^{\pi}(s,a_i) + \varepsilon(s,a_i)$$

where $\varepsilon(s,a_i)$ has zero mean and finite variance $\sigma_{s,a_i}^2$. Therefore, if $\hat{Q}^{\pi}(s,a_i)$ is an estimate of $Q^{\pi}(s,a_i)$ obtained by averaging $N$ observations of $Q^{\pi}_{\text{obs}}(s,a_i)$, then the mean and variance are given by:

$$E \left[ \hat{Q}^{\pi}(s,a_i) \right] = Q^{\pi}(s,a_i), \quad V \left[ \hat{Q}^{\pi}(s,a_i) \right] = \frac{\sigma_{s,a_i}^2}{N} \quad (4)$$

In addition, we assume that $Q^{\pi}_{\text{obs}}(s,a_i)$ are independently distributed. This is consistent with the MDP assumption.

3 Rate of Convergence Results

Before stating the convergence theorems, we define the following:

$$\sigma_{\text{max}}^2 = \max_{s \in S, a \in \{1, \ldots, M\}} \sigma_{s,a}^2, \quad \sigma_{\text{min}}^2 = \min_{s \in S, a \in \{1, \ldots, M\}} \sigma_{s,a}^2 \quad (5)$$

where $\sigma_{s,a}^2$ is defined in (4) and

$$C_{\text{min}} = \left[ \sum_s (d^{\pi}(s))^2 \sum_{i=1}^{M} \left( \frac{\partial \pi(s,a_i;\theta)}{\partial \theta} \right)^2 \right] \sigma_{\text{min}}^2$$

$$C_{\text{max}} = \left[ \sum_s (d^{\pi}(s))^2 \sum_{i=1}^{M} \left( \frac{\partial \pi(s,a_i;\theta)}{\partial \theta} \right)^2 \right] \sigma_{\text{max}}^2$$

3.1 Rate of Convergence of PIFA Algorithms

Consider the PIFA algorithm [5] which uses a basis function representation for estimated state action value function, $\hat{Q}^\pi$, of the following form:

$$\hat{Q}^\pi(s,a_i) = f^\pi_{a_i}(s) = \sum_{l=1}^{L} w_{a_i,l} \phi_{a_i,l}(s) \quad (7)$$

where $w_{a_i,l} \in \mathbb{R}$ are weights and $\phi_{a_i,l}(s)$ are basis functions defined in $s \in \mathbb{R}^D$. If the weights $w_{a_i,l}$ are chosen based using the observed $Q^{\pi}_{\text{obs}}(s,a_i)$, and the basis functions, $\phi_{a_i,l}(s)$, satisfy the conditions defined in [5, 4], then the performance gradient is given by:

$$\frac{\partial \rho}{\partial \theta} = \sum_s d^{\pi}(s) \sum_{i=1}^{M} \frac{\partial \pi(s,a_i;\theta)}{\partial \theta} f^\pi_{a_i}(s) \quad (8)$$

The following theorem establishes bounds on the rate of convergence for this representation of the performance gradient.
Theorem 1: Let \( \hat{\rho}_{\theta'} \) be an estimate of \( (8) \) obtained using the PIFA algorithm and the basis function representation \( (7) \). Then, given the assumptions defined in Section 2 and equations \( (5) \) and \( (6) \), the rate of convergence of a PIFA algorithm is bounded below and above by:

\[
C_{\min} \frac{ML}{N} \leq V \left[ \frac{\partial \rho}{\partial \theta_{F}} \right] \leq C_{\max} \frac{ML}{N} \tag{9}
\]

where \( L \) is the number of basis functions, \( M \) is the number of possible actions, and \( N \) is the number of independent estimates of the performance gradient.

Proof: See Appendix.

3.2 Rate of Convergence of Direct Sampling Algorithms

In the previous section, the observed \( Q_{\text{obs}}^{\pi} (s, a_i) \) are used to build a linear basis function representation of the state action value function, \( Q^{\pi} (s, a) \), which is in turn used to estimate the performance gradient. In this section we establish rate of convergence bounds for performance gradient estimates that directly use the observed \( Q_{\text{obs}}^{\pi} (s, a_i) \) without the intermediate step of building the FA representation. These bounds are established for the conditions \( b(s) = \frac{1}{M} \sum_{i} Q(s, a) \) and \( b(s) = 0 \) in \( (3) \).

Theorem 2: Let \( \hat{\rho}_{\theta} \) be an estimate of \( (3) \), be obtained using direct samples of \( Q^{\pi} \). Then, if \( b(s) = 0 \), and given the assumptions defined in Section 2 and equations \( (5) \) and \( (6) \), the rate of convergence of \( \hat{\rho}_{\theta} \) is bounded by:

\[
C_{\min} \frac{1}{N} \leq V \left[ \frac{\partial \rho}{\partial \theta} \right] \leq C_{\max} \frac{1}{N} \tag{10}
\]

where \( N \) is the number of independent estimates of the performance gradient. If \( b(s) \neq 0 \) is defined as:

\[
b(s) = \frac{1}{M} \sum_{j=1}^{M} Q^{\pi} (s, a_j) \tag{11}
\]

then the rate of convergence of the performance gradient \( \hat{\rho}_{\theta} \) is bounded by:

\[
C_{\min} \frac{1}{N} \left( 1 - \frac{1}{M} \right) \leq V \left[ \frac{\partial \rho}{\partial \theta_{\theta}} \right] \leq C_{\max} \frac{1}{N} \left( 1 - \frac{1}{M} \right) \tag{12}
\]

where \( M \) is the number of possible actions.

Proof: See Appendix.

Thus comparing \( (12) \) and \( (10) \) to \( (9) \) one can see that policy gradient algorithms such as PIFA which build FA representations of \( Q \) converge by a factor of \( O(ML) \) slower than algorithms which directly sample \( Q \). Furthermore, if the bias term is as defined in \( (11) \), the bounds on the variance are further reduced by \( O(1 - (1/M)) \). In the next section experimental evidence is given showing that these theoretical considerations can be used to improve the convergence properties of PGRL algorithms.

4 Experiments

The Simulated Environment: The experiments simulate an agent episodically interacting in a continuous two dimensional environment. The agent starts each episode in the same state \( s_i \), and executes a finite number of steps following a policy to a fixed goal state \( s_G \). The stochastic policy is defined by a finite set of Gaussians, each associated with a specific