INCREMENTAL REINFORCEMENT LEARNING

Abstract. Gradient methods fundamentally intertwine exploration and exploitation. For these techniques, determining if any gradient direction (in policy space) will improve the performance of the current policy can be arbitrarily difficult. To overcome these difficulties, an alternative performance measure which does not suffer from variance convergence problems is proposed and analyzed. Properly optimizing this alternative performance measure requires nonparametric techniques and can be done in a way that never noticeably decreases the average reward.

1. Introduction

In recent years, there has been a growing interest in direct policy-gradient methods for approximate planning in large Markov decision problems (MDP). Direct policy-gradient methods seek to find a good policy π among some restricted class of policies, by following the gradient of the future reward. Gradient methods are essentially on-policy and so intertwine the fundamental exploration-exploitation tradeoff. We argue that this entanglement could lead to highly inefficient policy improvement and suggest how to incorporate exploration into these methods.

Earlier work has argued that undirected policies could lead to exponentially inefficient learning [?]. For example, consider the simple MDP shown in Figure 1.1A (adapted from [?]). Under a policy that gives equal probability to all actions, the expected time to the goal from the left most state is $3(2^n - n - 1)$, and with $n = 50$, the expected time to the goal is about 10^{15}. Any gradient method initialized to this policy would have to run at least this long to obtain just one accurate estimate. The inherent problem is that in gradient methods, policy improvement at a state is tied to the current probability of visiting that state. Importance sampling methods have been formulated [?, ?] but these cannot salvage the situation when there is an exponentially low probability event with a significant contribution to the expectation of interest.

In value function methods, this problem is seen as a lack of exploration. Active exploration is necessary to improve the policy at states that are not visited often in order to efficiently propagate rarely visited state value information to commonly visited states (see [?, ?]). Dynamic programming also improves the policy at all states: unlikely states will exhibit policy improvement even though this improvement might have little impact on the average reward.

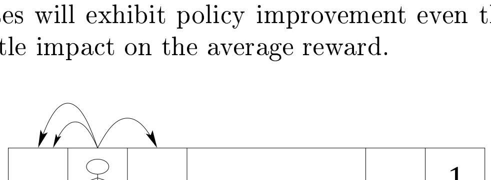

Figure 1.1. Example 1: Two actions move agent to the left and one actions moves agent to the right.

The average reward is what we ultimately seek to maximize. However, this performance measure has the fundamental problem: it is insensitive to performance improvement at highly unlikely states. Improvement at these states can lead to drastic policy improvement (as in figure 1.1). We formalize this by showing that it can be very difficult to improve a policy which has significantly sub-optimal performance. What is meant by difficult is that estimating gradient *direction* (in policy space) when the gradient *magnitude* is small can be exponentially difficult (in the number of states). Although being close to the optima implies the gradient has a small magnitude, a small magnitude does *not* imply being close to optima. The crucial problem is that the magnitude of the gradient can be small at plateaus far away from the optima.

- In section 2, the usual definition of the reinforcement learning problem is stated.
- In section 3, the problems with gradient methods are exhibited in more detail.
- In section 4, we discuss general incremental policy improvement.
- In section 6, an alternative performance measure is proposed which is sensitive to performance improvement at unlikely states (unlike the average reward). This alternative can in principle be improved with far fewer samples than the standard performance measure.
- In section 8, we discuss why richer function approximation schemes (such as non-parametric schemes) are necessary.

2. Preliminaries

First, let us state the standard formulation of a reinforcement learning problem. A reinforcement learning problem is a finite Markov decision process (MDP) with rewards which is a tuple (S, μ, A, R, P) where: S is finite set of states, μ is the starting state distribution, A is a finite set of actions, R is a reward function $R : S \times A \to [0, R_{max}]$, and P is the transition model in which p^a_{ij} is the transition probability from state i to j under action a. We only consider the more difficult case where the goal of the agent is to find a policy that maximizes the average reward. The agent's decision making procedure is characterized by a stochastic policy $\pi(a|s)$, which is the probability of taking action a in state s. We make the assumption that *every policy π is ergodic, ie* has a well-defined stationary state visitation frequency d_π.

Under this assumption, the *average reward* is $\sum_{s,a} d_\pi(s)\pi(a|s)R(s,a)$. Although in many situations, the goal of the agent is to maximize the average reward, we only consider the case of maximizing the γ discounted average reward. This approximation allows us to avoid mathematical difficulties while arriving at a similar solution for processes that "restart" on a reasonable timescale [?].

We define the state-action value as

$$Q_\pi(s,a) \equiv E_\pi\{(1-\gamma)\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)|s_0 = s, a_0 = a\} \tag{2.1}$$

and the value function as

$$J_\pi(s) \equiv E_{\pi(a'|s)}\{Q_\pi(s,a')\} \tag{2.2}$$

, where and s_t and a_t are the state and action at time t. Thus, the goal of the agent is to maximize the discounted average reward:

$$\eta(\pi) = E_\mu\{J_\pi(s)\}\,.$$

One last definition is required. The advantage is defined by:

$$A_\pi(s,a) \equiv Q_\pi(s,a) - J_\pi(s)\,. \tag{2.3}$$

We will sometimes suppress the π subscript when the policy is clear from context.

3. What is Wrong with Policy Gradients?

Let us consider the family of problems shown in figure 1.1 where an agent needs to reach some goal state. Policy gradient techniques will perform very poorly on these problems.

Theorem 3.1. *(Policy gradient is inefficient) There exist reinforcement learning problems with $n = |S|$ states which require an expected $\Omega(2^n)$ state transitions in order to accurately estimate the gradient direction.*

Proof. Consider the family of n state problems given in example 1.1 with table lookup policy placing equal probability on every action. The stationary distribution under this policy will satisfy:

$$\Pr(\text{state } i|i+1) = \frac{2}{3}$$

$$\Pr(\text{state } i+1|i) = \frac{1}{3}$$

$$\Pr(\text{state } i|i+1)\Pr(\text{state } i+1)$$
$$= \Pr(\text{state } i+1|i)\Pr(\text{state } i)$$

$$\Rightarrow \Pr(\text{state } i) = \frac{\Pr(\text{state } i+1)}{2}$$

In particular, this implies that the stationary probability of the goal state will be $\Theta(2^{-n})$. This, implies that the expected number of steps between visits to the goal state is $\Theta(2^n)$. All that remains is showing that a gradient estimate which does not visit the goal state will not be accurate. This can be done by noticing that the gradient estimate provided by any state sequence which does not visit the goal state will be identical to reinforcement learning where the goal transition reward is removed. This altered reinforcement learning problem has zero gradient and so the gradient direction must be arbitrary. □

Since Figure 1.1 is a very simple reinforcement learning problem, it is easy to imagine that many more complex problems have this simple problem embedded within the larger problem.

The situation is more generic and worse than theorem 3.1 might indicate. Let us give an additional example in which the situation is dire even for a two state example as shown in Figure 3.1B. Our example will use the common Gibbs table-lookup distribution, $\tilde{\Pi}_G = \{\pi_\theta : \pi(a|s) \propto \exp(\theta_{sa})\}$. The gradient of the discounted average reward (as shown later) is:

$$\partial_{sa}\eta(\pi) = \frac{1}{1-\gamma} d_\pi(s)\pi(a|s)A_\pi(s,a) \tag{3.1}$$

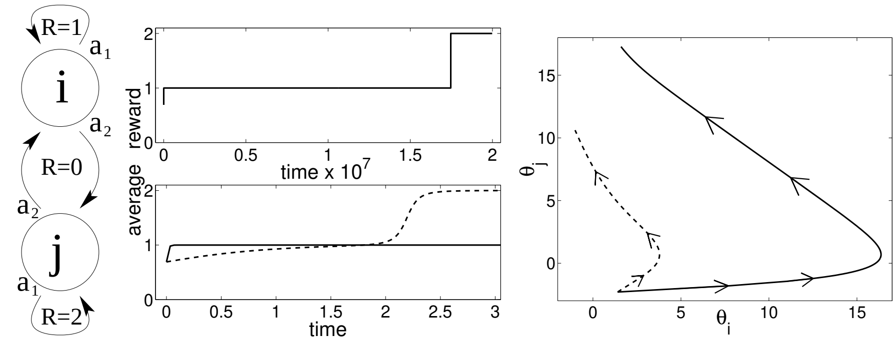

FIGURE 3.1. Example 2: A) See text. B top) The average reward vs. time (on a 10^7 scale) of a policy under standard gradient descent using the sigmoidal policy parameterization ($\pi(1; s, \theta_i) \propto \exp(\theta_i)/(1+\exp(\theta_i))$, with the initial conditions $\pi(i, 1) = .8$ and $\pi(j, 1) = .1$. B bottom) The average reward vs. time (unscaled) under standard gradient descent (solid line) and natural gradient descent (dashed line) for an early window of the above plot. D) Phase space plot for the standard gradient case (the solid line) and the natural gradient case (dashed line).

Increasing the chance of a self-loop at i decreases the stationary probability of j, which hinders the learning at state j. Under an initial policy that has the stationary distribution $\rho(i) = .8$ and $\rho(j) = .2$, learning at state i reduces the learning at state j leading to an an extremely flat plateau of improvement at 1 average reward shown in Figure 3.1C (see figure text). Figure 3.1D shows that this problem is so severe that $\rho(j)$ drops as low as 10^{-7} from it's initial probability of .2, which makes sampling the gradient difficult. Figure 3.1E shows where this occurs and gives a clearer picture (see figure text).

Sham, fill in the missing figures. -John

One question immediately arises: are we suffering from numerical error problems? The answer is no, as can be shown by a careful (and somewhat hairy) analysis of reinforcement learning with the Gibbs distribution. The analysis is given in section 7.

This example is remarkably discouraging for policy gradient approaches. We can easily make it even worse by using a few extra states in a chain as in figure 1.1.

4. Incremental Improvement

Gradient methods guarantee improvement due partially to the incremental nature of the algorithm although no *rate* of improvement guarantees are made. In contrast, dynamic programming with function approximation has only very weak performance guarantees (in terms of the worst case error) primarily due to the possibility of drastic policy changes. Before we analyze the problems with gradient methods, we first state the general conditions for incremental policy improvement.

Let's consider the incremental update rule:

$$\pi(a|s) \propto \pi(a|s)(1 + \alpha f(s, a)) \tag{4.1}$$

where f is some vector. Note that this update rule is to first order equal to the update rule $\pi(a|s) \propto \pi(a|s)\exp(\alpha f(s,a))$, which enforces positivity. Also note that using an infinite α returns a greedy policy based on f. If f is equal to the true advantages A, then the probability of actions with highest advantages increase the fastest.

Before we state the theorem, a few definitions are in order. Given a fixed starting state distribution $m_0(s_0)$, a fixed policy, $\pi(a|s)$, and a fixed Markov decision process, we can define the distribution at timestep t as:

$$d_{\pi,t}(s) = E_{\pi,\mu}\Pr(s_t = s|s_0)$$

We are most concerned with the γ discounted average distribution:

$$d_\pi(s) \equiv (1-\gamma)\sum_{t=0}^{\infty}\gamma^t d_{\pi,t}(s)$$

, which is analogous to the stationary distribution in the infinite time setting (*ie* as $\gamma \to 1$ then $d_\pi(s)$ tends to the stationary distribution) Also, we define the inner product for vectors in the state-action space as:

$$E_{d_\pi \pi} x x' = \sum_{sa} d_\pi(s)\pi(a|s)x(s,a)x'(s,a)$$

where $d_\pi(s)\pi(a|s)$ is the state-action distribution..

Theorem 4.1. *Let A be the true advantages. If the update rule is $\pi(a|s) \leftarrow \pi(a|s)(1+\alpha f(s,a))$, then the performance improvement of η is:*

$$\Delta\eta = \frac{\alpha}{1-\gamma}E_{d_\pi\pi}Af + O(\alpha^2)\,. \tag{4.2}$$

Proof. As shown in [?], the gradient of the discounted average reward from a starting state s_0 chosen according to μ is:

$$\nabla\eta(\pi) = \frac{1}{1-\gamma}E_{s\sim d_\pi(s)}\left[\sum_a Q_\pi(s,a)\nabla\pi(a|s)\right] \tag{4.3}$$

Let us consider the table-lookup Gibbs distribution, $\tilde{\Pi}_G = \{\pi_\theta : \pi(a|s) \propto \exp(\theta_{sa})\}$. Note that $\partial_{sa}\pi(a'|s') = \delta_{s's}\pi(a|s)(\delta_{a'a} - \pi(a'|s))$, where $\partial_{sa} \equiv \partial/\partial\theta_{sa}$. It follows that:

$$\begin{aligned}
\partial_{sa}\eta(\pi) &= \frac{1}{1-\gamma}d_\pi(s)\left[\sum_a Q_\pi(s,a)\pi(a|s)(\delta_{aa'} - \pi(a'|s))\right] && (4.4)\\
&= \frac{1}{1-\gamma}d_\pi(s)\pi(a|s)\left[Q_\pi(s,a) - E_{\pi(a'|s)}Q_\pi(s,a')\right] && (4.5)\\
&= \frac{1}{1-\gamma}d_\pi(s)\pi(a|s)A_\pi(s,a)\,. && (4.6)
\end{aligned}$$

For some change $\Delta\theta_{sa}$ to the parameters, the update rule is:

$$\pi(a|s,\theta_{t+1}) \leftarrow \pi(a|s,\theta_t)\exp(\Delta\theta_{sa})$$

, the resulting change to the performance is $\Delta\eta = \sum_{sa}\partial_{sa}\eta\Delta\theta_{sa} = \frac{1}{1-\gamma}\sum_{sa}d_\pi(s)\pi(a|s)A(s,a)\Delta\theta_{sa}$ (for small $\Delta\theta$).

The update rule for the table-lookup Gibbs is exactly the same as that used in equation 4.1, so we evaluate the change to the performance

$$\Delta\eta = \frac{1}{1-\gamma}\sum_{sa} d_\pi(s)\pi(a|s)A(s,a)\alpha f(s,a) = \frac{\alpha}{1-\gamma}E_{d\pi}Af$$

to first order. □

This theorem states that to obtain improvement all we need is the much weaker result that f must have a positive inner product with A (as opposed to strong and intractable conditions using the l_∞ norm).

Note that this theorem implies that if we define $\tilde{f}(s,a) \equiv f(s,a) - \sum_{a'}\pi(a'|s)f(s,a')$ then we get:

$$\Delta\eta = \alpha E_{d\pi}Q\tilde{f} + O(\alpha^2)\,.$$

This is important because it is difficult to estimate $E_{d\pi}fA$, since the advantages can't be obtained directly from Monte-Carlo estimates. However, we can directly estimate $E_{d\pi}Q\tilde{f}$ since the state-action values can be estimated from Monte-Carlo samples using the empirically observed discounted average reward.

Let us relate the previous theorem to policy gradient methods. For simplicity, let us consider the class of policies parameterized by one parameter, θ. The gradient (equation **??**) can be written in the form:

$$\begin{aligned}\frac{\partial\eta}{\partial\theta} &= \frac{1}{1-\gamma}E_{d\pi}\left[Q_\pi(s,a)\frac{1}{\pi(a|s)}\frac{\partial\pi(a|s)}{\partial\theta}\right] \\ &= \frac{1}{1-\gamma}E_{d\pi}\frac{\nabla\pi}{\pi}Q\,.\end{aligned}$$

The update rule is

$$\begin{aligned}\pi(a|s) &\propto \pi(a|s) + \frac{\partial\pi}{\partial\theta}\Delta\theta \\ &= \pi(a|s)(1+\Delta\theta\frac{1}{\pi(a|s)}\frac{\partial\pi(a|s)}{\partial\theta})\end{aligned}$$

and so $f = \frac{1}{\pi}\frac{\partial\pi}{\partial\theta}$. Thus we see that gradient descent is a special case of the previous result in which the direction followed is $\frac{1}{\pi}\frac{\partial\pi}{\partial\theta}$ or $-\frac{1}{\pi}\frac{\partial\pi}{\partial\theta}$ depending on the sign of $E_{d\pi}fQ$.

In order to improve the policy, it is *crucial* to obtain the correct sign of $E_{d\pi}\frac{1}{\pi}\frac{\partial\pi}{\partial\theta}Q$. As stated in the previous section, the variance in obtaining the correct sign can be large.

5. A Different Performance Measure

In this section, we show that the performance measure η is fundamentally inadequate. As discussed in section 3, the performance measure being used must be sensitive to improvements at unlikely states, where policy improvement might be necessary in order for the agent to achieve optimal payoff.

We desire an alternative performance measure that does not down weight advantages at unlikely states or unlikely actions. The performance measure we choose will *require* access to an exploration policy or simulator which can achieve a more "uniform" distribution over states. A natural candidate for a performance measure is to weight the improvement from all states and all actions more uniformly (rather than by μ), such as

$$\omega(\pi) = \sum_a \sum_s \mu_e(s,a) J_\pi(s).$$

where μ_e is some 'exploratory' starting *state-action* distribution. Note that $\omega = \eta$ for the special case when $\mu_e = \mu$. Under this performance measure, even if a state is not visited often (like the right most states in example 1), importance will be given to improving the future reward from this unlikely state. Since the policy that maximizes ω also simultaneously maximizes the future reward from every state, then maximizing ω also maximizes η.

Now let us consider for an arbitrary μ_e how difficult is it to improve ω using a direction given to us by some vector $f(s,a)$. As discussed in the last section, the choice of parameterization implicitly defines the vector $f(s,a)$. In the following subsections, we show that there exists a vector $f(s,a)$ for which determining if f or $-f$ will improve the average state values takes $O(1/\varepsilon)$ samples when the state values are not $O(\sqrt{\varepsilon})$ near optimal. The same statement cannot be made for policy gradients of η. The subsections are outlined as follows:

- in 5.1, we bound the sample size necessary to improve ω
- in 5.2, we argue why using μ_e different from μ is important
- in 5.3, we discuss the implications of improving ω rather than η

5.1. **The Difficulty of Improving ω.** Given a direction f how hard is it to determine if this direction will improve the policy? We must first consider how information about the direction of f can be found. Fundamentally, the only process available to us is running the MDP of the reinforcement learning problem. Given certain assumptions 8, it is reasonable to suppose that we can calculate a quantity $\hat{Q}(s,a)$ which is a nearly unbiased estimate of $Q(s,a)$. Note that this value is relative to the current value of the policy. Given that all rewards are in some interval $[0, R_{max}]$, these estimates will have some finite variance, $\sigma(s,a)$. Our goal is now to prove a statement about the number of samples required in order to estimate if a particular direction in state-action space given by $f(s,a)$ is in fact the "right" direction. In order to do this properly, we must first define a constrained f:

$$\tilde{f}(s,a) \equiv f(s,a) - \sum_{a'} \pi(a'|s) f(s,a')$$

With this definition, we have the property that for all states s, $E_{\pi(a|s)}\tilde{f}(s,a) = 0$. Without loss of generality we also assume that $\tilde{f}$ is of unit length, *ie* $\left\langle \tilde{f}, \tilde{f} \right\rangle = 1$. Also, let us define the analogous discounted future distribution of states as:

$$\tilde{d_\pi}(s) \equiv (1-\gamma) \sum_{t=0}^{\infty} \gamma^t E_{\pi,\mu_e(s_0)} \Pr(s_t = s | s_0).$$

Under theorem 4.1, ω will improve when the policy π is perturbed toward f if $E_\nu Q, \tilde{f} > 0$, where $\nu(s,a)$ is an arbitrary distribution over state-action choices. For example, this distribution might be derived from $\mu_e(s)$.

Theorem 5.1. *Assume that samples s,a are drawn from some distribution $\nu(s,a)$ and that $E_\nu \bar{f}^2 = 1$. For each sample, we are given an unbiased estimate $\hat{Q}(s,a)$,* *of $Q(s,a)$ with variance $\sigma^2(s,a)$. Let $\sigma_{min} = \min_{s,a}\sigma(s,a)$ and assume all rewards satisfy $R \in [0, R_{max}]$. Then the estimate $\hat{Q}(s,a)\tilde{f}(s,a)$ of $E_\nu Af$ has variance bounded by:*

$$\sigma^2_{min} - (E_\nu A\tilde{f})^2 \leq \Sigma^2 \leq \frac{R^2_{max}}{4}\,.$$

Proof. Since the $\hat{Q}$ are unbiased, the mean μ of the estimate $\hat{Q}(s,a)\tilde{f}(s,a)$ is also unbiased. The variance of this estimate is:

$$\Sigma^2 = E_\nu \tilde{f}^2\hat{Q}^2 - \mu^2 \tag{5.1}$$

$$= \sum_{s,a}\nu(s,a)\int dP(\hat{Q}(s,a))\hat{Q}^2(s,a)\tilde{f}^2(s,a) - \mu^2 \tag{5.2}$$

$$= \sum_{s,a}\nu(s,a)\tilde{f}^2(Q^2(s,a)+\sigma^2(s,a)) - \mu^2\,. \tag{5.3}$$

We can lower bound σ as follows:

$$\begin{aligned}\Sigma^2 &\geq \sum_{s,a}\nu(s,a)\tilde{f}^2\sigma^2(s,a) - \mu^2\\ &\geq \sigma^2_{\min}\sum_{s,a}\nu(s,a)\tilde{f}^2(s,a) - \mu^2\\ &= \sigma^2_{\min} - \mu^2\end{aligned}$$

where have used the unit length of $\tilde{f}$. We can upper bound Σ^2 by noting that each sample $\hat{Q}(s,a)$ is drawn from the interval $[0, R_{\max}]$ and $E\hat{Q}(s,a) = Q(s,a)$. This implies that the maximum variance for any state-action is $\frac{R^2_{\max}}{4}$. Therefore, the expected variance is bounded by $\frac{R^2_{\max}}{4}$. □

In order to determine if the direction f will improve the policy, we have to determine with some certainty the correct sign of the mean $\mu = E_\nu A\tilde{f}$. The difficulty of this is determined by how much larger the "signal" μ is than the "noise" Σ. If the ratio Σ/μ is large, then determining the sign $E_\nu A\tilde{f}$ will require many samples.

Corollary 5.2. *Under the assumptions and sampling model of theorem 5.1, then for any f, the noise to signal is bounded by*

$$\frac{\Sigma^2}{\mu^2} \geq \frac{\sigma^2_{min}}{E_\nu A^2} - 1\,.$$

Proof. The $\tilde{f}$ that maximizes $E_\nu A\tilde{f}$ also minimizes the lower bound on Σ^2 (subject to the normalization constraint on $\tilde{f}$). This occurs when $\tilde{f} = \frac{A}{\sqrt{E_\nu A^2}}$. Substitution leads to the result. □

Lets consider the special case of $\omega = \eta$ when $\mu_e = \mu$. If all the large advantages are at states that are visited infrequently then $E_\nu A^2$ will be small. The corollary shows that it is difficult to determine if any direction will improve the policy. Note that this is what occurs in the examples of the previous section. In the two state example, at the plateau, state j is visited highly infrequently, and so $E_\nu A^2$ is small, eventhough $E_\nu A^2$ was large to begin with.

5.2. **Why is ω better?** Let us bound the number of samples $\hat{Q}(s,a)$ needed to determine if f will improve. The variance will decrease like $\frac{1}{k}$ where k is the number of samples.

Corollary 5.3. *Under the assumptions and sampling model of theorem 5.1, then the estimate* $X = \frac{1}{k}\sum_{i=1}^{k}\hat{Q}(s,a)\tilde{f}(s,a)$ *has the*

$$\Pr(sign(X) = sign(E_\nu Af)) \le e^{-\frac{2k(E_\nu Af)^2}{R_{max}^2}}$$

Proof. Hoeffding's inequality for k independent identical random variables, $X \in [0, R_{\max}]$ states:

$$\Pr\left(EX - \frac{1}{k}\sum_i X_i > \epsilon\right) \le e^{-\frac{2k\epsilon^2}{R_{\max}^2}}$$

setting $\epsilon = EX = E_\nu Af$ gives the result. □

We know that an advantage of size A will imply a bias of size at least $\mu = (1-\gamma)\frac{A}{n}$. Here n is the number of state-actions and we are assuming a uniform distribution over state-actions. This will, in turn, imply that at most $k = \frac{n^2}{A^2(1-\gamma)^2}$ samples are required in order to estimate the "right" direction with high confidence. This quantity is polynomial in all parameters and so given that a significant advantage exists anywhere, it convergences far faster than policy gradient estimates.

As suggested by corollary 5.2, if $E_\nu A^2$ is small then determining if any f will improve the policy will be difficult. If the policy has near optimal performance, then all the advantages are small and we expect improvement to be difficult regardless of the performance measure (*ie* μ_e) being used. This is the common situation that close to a maxima it is difficult to determine if any direction will improve the function due to local flatness. The ideal situation is that the estimates only have high variance when the policy has near optimal performance. Unfortunately, as demonstrated by our two examples, this is not the case for η. The important question we now address is "How difficult is it to determine if f improves a sub-optimal policy?"

First, let us provide a theorem as to how far the policy is from optimal in terms of how large $E_\nu A^2$ is.

Theorem 5.4. *Let $J^*(s)$ be the maximum value of $J(s)$, $\nu_0 = \min_{s,a}\nu(s,a)$, and $A_\infty = \max_{s,a} A(s,a)$. Then*

$$|J(s) - J^*(s)| \le \frac{A_\infty}{1-\gamma} \le \frac{1}{1-\gamma}\sqrt{\frac{E_\nu A^2}{\nu_0}}. \tag{5.4}$$

Proof. Let $A_\infty = \max_{s,a} A(s,a)$. By definition of the advantages, if we deviate from the current policy at any step we can gain at most A_∞ future reward. We can deviate at every steps, and by induction, $|J(s) - J^*(s)| < A_\infty/(1-\gamma)$ (*Cite book, or prove it. -John).* The largest possible A_∞ for a given $E_\nu A^2$ satisfies $E_\nu A^2 = \nu_0 A_\infty^2$. □

Note that this implies for any distribution μ and for $\omega(\pi) = E_\mu\{J_\pi(s)\}$,

$$(\omega - \omega^*)^2 \le \frac{E_\nu A^2}{\nu_0(1-\gamma)^2} \tag{5.5}$$

where ν is arbitrary (not necessarily related to μ).

This bound elucidates the problems with using μ as the starting distribution. Let us consider the case in which $E_\nu A^2$ is small, so determining if f or $-f$ improves the policy will be difficult. If ν_0 is small, then A_∞ could be large even if $E_\nu A^2$ is small. Thus, η could be far from η^*, since each $J(s)$ is significantly sub-optimal. Essentially, the large advantages are hidden to the agent due to the low probability of visiting the state-actions with large advantages. This makes policy improvement difficult regardless of the choice of f. Unfortunately, the presence of large advantages leaves open the possibility that the policy is significantly sub-optimal.

Instead, let us consider a μ_e that does not allow ν_0 to be extremely small. This will force each $J(s)$ to be close to optimal if $E_\nu A^2$ is small. The number of samples must depend on the direction f being followed, and there are directions for which determining whether f or $-f$ improves the policy has high variance. For example, directions which "alias" important distinctions between the states can have high variance.

In summary, for a near uniform μ_e, we know that the distance from optimal is constrained to satisfy theorem 5.4, and if we are far from optimal there exists a direction $f(s,a)$ which can be confirmed as "good" with $k = O\left(\frac{R^2_{\max}}{(E_\nu A f)^2}\right)$ samples.

5.3. **What about η?** *I weakened this section to essentially state that, in principle, it is possible to 'usually' improve η when ω improves. -Sham*

What has been lost by improving ω instead of η? Our real goal *is* improving η. An important question is: "Can η decrease as we search for this near-optimal policy?" In general, the answer is "yes" because we are only guaranteed to increase ω. To see this, consider $f(s,a) = A(s,a)$ on all states visited often under $d_\pi(s)$ and $f(s,a) = -A(s,a)$ on all other states. This will force $E_\nu f A > 0$ under the exploitation policy. However, if the state distribution from μ_e has high probability at states where $f(s,a) = -A(s,a)$ then η can decrease when we optimize according to ω. For a sufficiently flexible space of $f(s,a)$, the ultimate policy will be near-optimal, but increasing ω does not always imply η improves. There *do* exist $f(s,a)$ for which optimizing $\omega(\pi)$ always improves $\eta(\pi)$. To see this, simply make $f(s,a) = A(s,a)$ everywhere.

In general, determining if any f or $-f$ improves η at every step is impractical, as shown previously. *This is why we considered ω to begin with.* However, for all other cases we would like to determine if f improves both ω and η with $O(1/\varepsilon)$ samples, if the policy is not $O(\sqrt{\varepsilon})$ near optimal.

I'm losing the thread again here. The word 'oracle' confuses me and I think you work with strange "derived" distributions. I'll let you go through it again and see if you can make sense of it. -John

Let ν and $\tilde{\nu}$ be the distributions implied by μ and μ_e respectively and let us consider the case in which we are not guaranteed to be $\frac{1}{1-\gamma}\sqrt{\frac{\varepsilon}{\nu_0}}$ near optimal policy. Theorem **??** states that there exists an oracle such that ω can be improved with $O(1/\varepsilon)$ samples. As far as η is concerned, let us consider the non-problematic case in which the agent also visits states with large advantages under μ, *ie* $E_\nu A^2 > \varepsilon$. Also by theorem **??**, there exist a different oracle such that improvement of η requires $O(1/\varepsilon)$ samples. Further, there are $\tilde{f}$ that jointly satisfy the conditions of each oracle, $(E_\nu \tilde{f} A)^2 > \varepsilon$ and $(E_{\tilde{\nu}} \tilde{f} A)^2 > \varepsilon$ (when not in the problematic case for η). For instance, $\tilde{f} \propto A$ or any $\tilde{f}$ sufficiently parallel will satisfy both of these conditions. For these $\tilde{f}$, then we can determine if both η and ω will improve with $O(1/\varepsilon)$ samples.

The big picture is that we have lost very little by improving ω, if we can find these good directions f. First, we can improve $J(s)$ on average, under μ_e, with $O(1/\varepsilon)$ samples until η is guaranteed to be near-optimal. Second, also with $O(1/\varepsilon)$ samples, we can determine if η will improve as well, in the non-problematic cases. Thus all that we have lost in principle is that improvement might not be possible for η when the advantages are at unlikely states, which is unavoidable if we will only collect $O(1/\varepsilon)$ samples. However, remember that optimizing ω will also ultimately optimize η. We now discuss how to find these good directions f.

6. The Need for Non-Parametric Function Approximation

There is no reason to believe that a "good" choice of the vector $f(s,a)$ can be made without significant prior knowledge about the form of the reinforcement learning problem. Given some number of samples drawn from $\nu(s,a)$, we can test the value of a particular vector $f(s,a)$. The vector f or $-f$ may satisfy: $E_\nu fA > \epsilon$, in which case we can improve the policy with high probability. Or, it may be that the expectation is to small and we will be unable to determine whether or not the f is a beneficial perturbation given the (necessarily finite) samples from $\nu(s,a)$. This last case is by far the most common, and so it is necessary that we test many different f with the same set of samples from $\nu(s,a)$. This necessity creates the possibility of overfitting: we might find an f which works *only* on our sample set and not on future draws from $\nu(s,a)$.

There are two approaches to avoiding overfitting. One is to work with a finitely parameterized set F of vectors $f(s,a)$, and then simply use enough samples from $\nu(s,a)$ that we can guarantee (via VC bounds or other approaches) that the empirical estimate of $E_\nu fA$ must be strongly related to the actual value of $E_\nu fA$. The disadvantage of this approach is that it might not work. In particular, it is unclear (without extra prior knowledge) that we can guarantee there exists $f \in F$ with a large dot product ($E_\nu fA$). This is exactly what we wish to avoid by shifting from optimizing η to ω.

The other approach is "non-parametric" which is a great misnomer. In a non-parametric approach the number of parameters grows with the number of samples, but always remains small enough that we can guarantee that the empirical estimate of $E_\nu fA$ remains close to its real value. This is the approach that we advocate here: choose an $f(s,a)$ by sampling from $\nu(s,a)$ and testing until an $f(s,a)$ is found which improves ω with high probability. In the process of discovering $f(s,a)$, increase the space of functions as more samples are acquired. We can *always* check to see if a "good" $f(s,a)$ has been learned by simply acquiring a few extra independent samples from $\nu(s,a)$ and testing to see that $f(s,a)$ behaves well on the extra samples. If the test fails, then we have more samples from which to learn a new f.

Our suggested technique then does the following:

(1) Sample k samples from $\nu(s,a)$ and calculate $\hat{Q}(s,a)$ for each sample.
(2) Find an $f(s,a) \in F(k)$ for which the empirical estimate of $\frac{1}{k}\sum_{i=1}^{k} \bar{f}(s,a)Q(s,a)$ is largest.
(3) Verify that $f(s,a)$ does well with l new independent samples drawn from $\nu(s,a)$ (or other sample complexity approaches if applicable). If no, then go to step 2 with an extra l samples.

(4) Alter the policy by $\pi(a|s) \propto \pi(a|s) + \alpha f(s,a)$ and go to step 1.

This approach will always result in improvement according to the metric ω and so it will always result in improvement given that significant advantages exist.

Another advantage of the nonparametric approach is that we can avoid trading off between improvement of ω and improvement of η. Samples from η are easier to acquire so we could easily acquire an addition k samples from η and then find the optimal f for the empirical distribution which is a mix of the $\nu(s,a)$ and $d_\pi(s)\pi(a|s)$. If the choice of f makes optimization on the $d_\pi(s)\pi(a|s)$ samples essentially orthogonal to optimization on the $\nu(s,a)$ samples, then we expect that $f(s,a)$ will almost never decrease the value η.

(I don't quite know know how to work this into the story line. -John)

One other

A "good" set of f directions might satisfy the following criterion:

$$E_{\tilde{d}\pi}(f-A)^2 + E_{d\pi}(f-A)^2 \le \varepsilon^2 \,. \tag{6.1}$$

These f will always improve ω when not $O(\sqrt{\varepsilon})$ near optimal. Further, these f can simultaneously improve η unless the policy is near optimal.

References

7. Appendix

The goal of this section is constructing a proof that policy gradient will *eventually* work on a Gibbs table lookup distribution. Note that the time scale of "eventually" can be remarkably long.

The Gibbs table-lookup distribution is $\tilde{\Pi}_G = \{\pi_\theta : \pi(a|s) \propto \exp(\theta_{sa})\}$. Using the notation $\partial_{sa} \equiv \partial/\partial\theta_{sa}$,

$$\partial_{sa}\eta(\pi) = d_\pi(s)\pi(a|s)(Q_\pi(s,a) - J_\pi(s)) = d_\pi(s)\pi(a|s)A_\pi(s,a) \tag{7.1}$$

Note that for any deterministic policy (*ie* $\forall s\, \exists a$ s.t. $\theta_{sa} = \infty$) the gradient is zero. Since the gradient is 0 at all points at infinity, this suggests the dismal possibility that the gradient descent could lead to arbitrary deterministic policy. It is straight forward to show that if the stationary distribution, of the asymptotic policy, is nonzero for all states then this asymptotic policy is optimal. However, the nontrivial question is can $\partial_{sa}\eta_\pi$ tend to 0 due to $d(s)$ tending to 0 in a manner such that such that the optimal policy is not achieved. Note that the gradient $\partial_{sa}\eta(\pi)$ is zero if either the advantage is zero or if $d(s)$ is zero. The following theorem states that this can never occur, and the key to proof is dealing with this case (where learning to one state could stop due to a vanishing stationary probability). We only consider the infinitesimal learning rate case where $\theta_t = \theta_0 + \int_0^t \nabla\eta^{\pi_t}\, dt$.

Theorem 7.1. *Assume that π_∞ is the single limiting policy corresponding to $\lim_{t\to\infty}\theta_t$. Then π_∞ is globally optimal.*

Proof. Let $S^* = \{s | \exists\pi \text{ s.t. } \rho_\pi(s) \ne 0\}$. Our proof that π_∞ globally maximizes the average reward consists of showing that for all $s \in S^*$, $J^{\pi_\infty}(s) = \max_a Q^{\pi_\infty}(s,a)$, *ie* π_∞ partially satisfies the Bellman equations. When clear, the subscript on π_t is dropped, *eg* $\rho^{\pi_\infty} = \rho^\infty$. Assume for now that as $t \to \infty$ then for all $s \in S^*$ and actions a either $\pi_t(a|s) \to 0$ or $A^t(s,a) \to 0$. For all $\tilde{a} \in \text{argmax}_a Q^{\pi_\infty}(s,a)$, $\partial_{sa}\eta^t$ is positive for sufficiently large t, so $\pi_\infty(a|s) \ne 0$ and $A^\infty(s,a) = 0$. Therefore, π_∞ satisfies the Bellman equations for $s \in S^*$. It follows that if $\tilde{a} \notin \text{argmax}_a Q^{\pi_\infty}(s,a)$

then $\pi_\infty(\tilde{a}, s) = 0$, and so if t is sufficiently large $\partial_{\tilde{s}\tilde{a}}\eta^t < 0$, else there must exist a partial extrema for some $\tilde{t} \geq t$.

It is straightforward to show that $\lim_{t\to\infty} \partial_{sa}\eta^t = 0$, and so as $t \to \infty$ either $\rho^t \to 0$, $\pi_t(a|s) \to 0$, or $A^t(s,a) \to 0$. We now show that even if if $\rho^\infty \neq 0$, then $\pi_\infty(a|s) = 0$ or $A^\infty(s,a) = 0$, which completes the proof. Assume that for some s that $\rho^\infty(s) = 0$. By stationarity, $\rho^t(s) = \sum_{s_0} P(s_\tau = s|s_0)\rho^t(s_0)$, and so for any trajectory $s_0, a_0, s_1, a_1 \ldots s$, $\rho^t(s) \geq \rho^t(s_0) p^{a_0}_{s_0 s_1} \pi_t(a_0|s_0) \ldots p^{a_{\tau-1}}_{s_{\tau-1}s} \pi_t(a_{\tau-1}|s_{\tau-1})$. For every τ-length trajectory with $p^{a_i}_{s_i s_{i+1}} \neq 0$ and $\rho^\infty(s_0) \neq 0$, there must exist at least one s_i, a_i such that $\pi_\infty(a_i|s_i) = 0$, else $\rho^\infty(s) \neq 0$ or $s \notin S^*$. It can be shown that there exist some $\tilde{s}$, $\tilde{a}$ and positive constant $\tilde{\alpha}$ such that $\forall t\, \rho^t(s) \geq \tilde{\alpha}\rho^t(\tilde{s})\pi_t(\tilde{a}|\tilde{s})$, $\rho^\infty(\tilde{s}) \neq 0$, and $\pi_\infty(\tilde{a}|\tilde{s}) = 0$ (using a trajectory such that $\pi_\infty(a_i|s_i) = 0$ only for $i = 0$). Assume that $\pi_\infty(s) \neq 0$ and $A^\infty(s) \neq 0$ (else the conjecture is true), then it follows from equation 7.1 and the previous inequality that for each a, $|\int_0^\infty \partial_{sa}\eta^t dt| \geq \alpha |\int_0^\infty \partial_{\tilde{s}\tilde{a}}\eta^t dt|$ for some positive constant α. However, $\pi_\infty(\tilde{s}) = 0$ forces $|\int_0^\infty \partial_{\tilde{s}\tilde{a}}\eta^t dt|$ to be ∞, and so $|\int_0^\infty \partial_{sa}\eta^t dt|$ is ∞, contradicting $\pi_\infty(s) \neq 0$ and $A^\infty(s) \neq 0$. $\square$

8. How to Monte Carlo

In order to determine which direction is the “right” direction, we need to know how to estimate this right direction. This is a subtle problem when you must (in general) sample from the underlying reinforced MDP forever in order to acquire an unbiased estimate of the discounted future reward, $\hat{Q}(s,a)$.

Intuitively, if we run the Markov process for a time long compared to the discount γ, we will obtain an estimate $\hat{Q}(s,a)$ which is “nearly” unbiased, simply due to the fact that the contribution to the reward from future actions will necessarily be small.