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 function approximation 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.

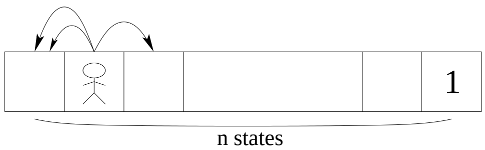

FIGURE 1.1. Example 1: Two actions move agent to the left and one actions moves agent to the right.

The average reward is what we ultimately seek to maximize. However, this performance measure has a 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 detail 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 5, 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 6, 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. 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 can not be accurate. Define a new reinforcement learning problem which is identical to 1.1 except with no reward on the goal transition. The gradient estimate provided by any state sequence which does not visit the goal state will be identical for both reinforcement learning problems. However, the altered reinforcement learning problem has zero gradient and so the gradient direction returned 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:

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

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

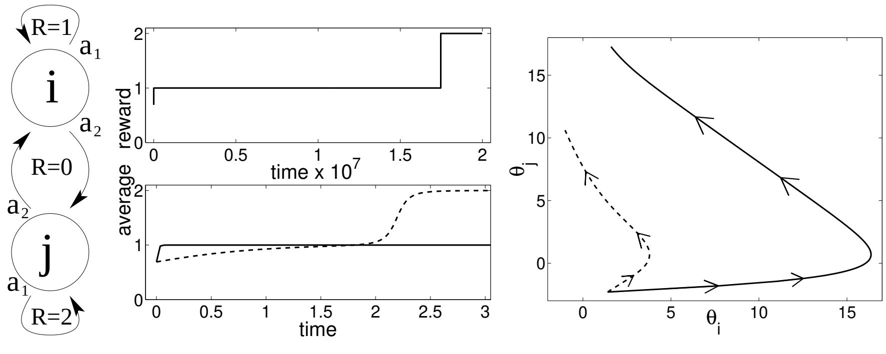

FIGURE 3.1. Example 2: A) See text. B top) The average reward vs. time (on a 10^7 scale) of a policy under standard gradient descent using the sigmoidal policy parameterization ($\pi(1; s, \theta_i) \propto \exp(\theta_i)/(1 + \exp(\theta_i))$), with the initial conditions $\pi(i, 1) = .8$ and $\pi(j, 1) = .1$. 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).

Sham, I don't understand the notation in this caption. What is s? What is "1"? -John

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 8.

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. CONDITIONS FOR 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 in the space of state-actions. 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 μ, a fixed policy, $\pi(a|s)$, and a fixed Markov decision process, we can define the distribution at timestep t as:

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

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, note that the expectation:

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

can be seen as an inner product for vectors x and x' in the state-action space under the state-action distribution $d_\pi(s)\pi(a|s)$.

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

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

where μ_e is some 'exploratory' starting *state-action* distribution. Note that $\omega = \eta$ for the special case when $\mu_e = \mu(s)\pi(a|s)$. 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 it is 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 $O(1/\varepsilon^2)$ samples when the state values are not ε-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 a 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 (see section 9), it is reasonable to suppose that we can calculate a quantity $\hat{Q}_\pi(s,a)$ which is a nearly unbiased estimate of $Q_\pi(s,a)$. Given that all rewards are in some interval $[0, R_{\max}]$, these estimates will have some finite variance, $\sigma(s,a) \leq \frac{R_{\max}^2}{4}$. 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:

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

Under theorem 4.1, ω will improve when the policy π is perturbed toward f if $E_\nu Q \tilde{f} > 0$. Here $\nu(s,a)$ is an arbitrary distribution over state-actions.

Sham this last statement is _NOT_ true right now and needs to be fixed, either by generalizing 4.1 (preferred) or weakening the statement. I'll let you handle it.

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 $E_\nu A f$ has variance bounded by:*

$$\sigma_{min}^2 - (E_\nu A\tilde{f})^2 \le \Sigma^2 \le \frac{R_{max}^2}{4}\,.$$

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:

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

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 $\frac{R_{\max}^2}{4}$. Therefore, the expected variance is bounded by $\frac{R_{\max}^2}{4}$. □

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 $E_\nu A\bar{f}$ 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*

$$\frac{\Sigma^2}{\mu^2} \ge \frac{\sigma_{min}^2}{E_\nu A^2} - 1\,.$$

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.

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*

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

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

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

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

Note that if $E_\nu Af$ is large then few samples are needed. In particular, the number of samples required to achieve a fixed error rate is $O\left(\frac{R_{\max}^2}{(E_\nu Af)^2}\right)$. This suggests that a uniform distribution is sensible since it will allow $E_\nu Af$ to have a significant value as long as any advantage is large. The next section formalized this need for 'exploration'.

I don't understand or agree with the comments that were previously here. The previous statement 'polynomial in all the parameters' was very weak, since the scalar A was one of the parameters (which wasn't defined). One of these parameters cannot be A, since A could scale with n (as in example 1). Policy gradients might converge 'slower' (it's a tricky argument to make at this point) but at least we are improving what we are interested in. It is a bit rude at this point to say that this different quantity is easy to improve, since, after all, it isn't the quantity we are interested in. The next section formalizes a criticism. -Sham

I don't see why A is a scalar. It looks like a vector to me. I think it is worth pointing out that the number of samples behaves well as a function of $\frac{1}{E_\nu Af}$ *and* R_{max}. *See if the current statement is sensible. -John*

5.2. **Why is exploration necessary?** 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?"

We measure the degree of sub-optimality by:

$$\varepsilon = \max_s |J(s) - J^*(s)|$$

which is the standard ∞-measure used in dynamic programming (see [**?**]). It follows that $|\eta - \eta^*| \leq \varepsilon$. Clearly, increasing an average of $J(s)$, under a distribution that has nonzero weight for every state, will eventually decrease this measure , and force the policy to be at least ε close to optimal.

The following theorem addresses the difficulty of improving ω for an ε sub-optimal policy. Let θ be the angle between the vectors A and f in state action space. Then the inner product satisfies $E_\nu Af = \sqrt{(E_\nu A^2)(E_\nu f^2)}\cos\theta$.

Theorem 5.4. *Let* $\cos\theta = \frac{E_\nu Af}{\sqrt{(E_\nu A^2)(E_\nu f^2)}}$, $\varepsilon = \max_s |J(s) - J^*(s)|$, *and* $\nu_0 = \min_{s,a}\nu(s,a)$. *Under the assumptions and sampling model of theorem 5.1, then the estimate* $X = \frac{1}{k}\sum_{i=1}^{k}\hat{Q}(s,a)\tilde{f}(s,a)$ *satisfies:*

$$\Pr(sign(X) = sign(E_\nu Af)) \le \exp(-k\varepsilon^2\nu_0\frac{2(1-\gamma)^2\cos^2\theta}{R_{max}^2})$$

Proof. First, let us bound ε in terms of how large $E_\nu A^2$ is. Let $A_\infty = \max_{s,a} A(s,a)$. By definition of the advantages, if we deviate from the current policy at any state s by taking action a, we gain, on average, $A(s,a)$ future reward. The change in $J_\pi(s)$ when we switch to any new policy π' is $E_{\mu_e,\pi'}(\sum_t \gamma^t A(s_t,a_t))$, since the new policy deviates from the old at every time step. Each term in this sum is at most A_∞, so for any policy $J_{\pi'}(s) - J_\pi(s) \le \frac{A_\infty}{1-\gamma}$. Hence,

$$\varepsilon \le \frac{A_\infty}{1-\gamma}. \tag{5.4}$$

The largest possible A_∞ for a given $E_\nu A^2$ satisfies $E_\nu A^2 = \nu_0 A_\infty^2$, so $A_\infty \le \sqrt{\frac{E_\nu A^2}{\nu_0}}$ and

$$\varepsilon \le \frac{1}{1-\gamma}\sqrt{\frac{E_\nu A^2}{\nu_0}}. \tag{5.5}$$

By definition of θ, $(E_\nu Af)^2 = E_v(A^2)E_\nu(f^2)\cos^2\theta \ge (1-\gamma)^2\nu_0\varepsilon^2\cos^2\theta$. (for normalized f), and the result follows using corollary 5.3. □

Thus, when the policy is ε-near optimal, we need at most $O\left(\frac{1}{\varepsilon^2\nu_0}\right)$ samples to estimate the gradient. If we use a μ_e that implies a uniform state-action distribution, then we need at most $O\left(\frac{n}{\varepsilon^2}\right)$ (where n is the number of state-actions).

The same cannot be said for policy gradients since ν_0 can be arbitrarily small, implying that arbitrarily many samples might be necessary. Essentially, ε can be large which implies that there must be large advantages (by equation 5.4), but the large advantages could be hidden from the agent in the sense that $E_\nu(A^2)$ can be small due to the skewed state action distribution. Thus, for large ε, the signal to noise could be large (by corollary 5.2) implying many more samples are necessary.

Although we ultimately seek to improve η, the appropriate performance measure to use is one which weights $J(s)$ more uniformly. This stems from the fact that if $|J(s) - J^*(s)|$ is large for just one state, then all other states stand to gain significant improvement. Intuitively, using ω is analogous to an incremental version of policy iteration.

5.3. **What about η?** What has been lost by improving ω instead of η? Our real goal *is* improving η. Note that the endpoint of improving ω will be a policy which is optimal for criteria 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 a f which improves the policy under the the distribution $d_\pi(s)$ yet $-f$ improves the policy under $\tilde{d_\pi}(s)$. Such f can exist when the initial distributions μ and μ_e differ significantly. For these f, optimizing according to ω will decrease η. There *always* exist $f(s,a)$ for which optimizing $\omega(\pi)$ improves $\eta(\pi)$: consider $f(s,a) = A(s,a)$ 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. Let ν be the distributions implied by μ, *ie* $d_\pi \pi$. The problematic cases for improving f are when $E_\nu(A^2)$ is small. The number of samples required to determine if f improves is $O\left(\frac{1}{(E_\nu Af)^2}\right)$. Using the angle between A and f determined by ν, the number of samples required is $O\left(\frac{1}{\cos^2\theta E_\nu A^2}\right)$. Thus, if $E_\nu(A^2)$ is not extremely small then it is possible to determine if f or $-f$ improves η.

We now focus on the important $\cos\theta$ dependence, and show how it is possible to improve both ω and η in nonproblematic case.

6. Policy Improvement

The question now is how to choose an f to improve the policy. An obvious method is to chose some f randomly and try to improve the policy the policy with it. The infamous 'curse of dimensionality' harms this approach in general. Recall from theorem 5.4 that $O\left(\frac{1}{\varepsilon^2\cos^2\theta}\right)$ samples allow us to determine the improvement direction. The problem is that a random f will have exponentially small overlap in the size of the state-action space with the vector A. Thus, $\cos\theta$ will be exponentially small and many samples might be necessary.

This last statement seems false. Assume that you have a normalized vector x and pick another random normalized vector x'. $Exx' = \Theta\left(\frac{1}{\sqrt{n}}\right)$. -John

There are two alternatives to improve the policy, which we know discuss. There is the brute force method of just optimizing ω. The second and more powerful method is to use function approximation to find an f that is aligned with A.

6.1. The Brute Force Method.

We could just try to use a gradient method to improve ω with some parameterized class of policies Π_θ. There are two problems with doing this. The first is that the parameterized class of polices must be tailored such that f always overlaps with A. In many problems, such information might not be available. The second and more serious problem is that Π_θ might not be a rich enough set such that a near-optimal policy is contained, so when θ converges to a local optimum η might still be small. In order to achieve near-optimal performance, the "size" of the set Π_θ must be increased. For instance, if the class is parameterized by a neural network and θ converges to point where η is insufficiently small, more hidden units must be added.

The advantage of this method is that a compact representation of the policy can be kept.

I think this quite a big concern. With the function approximation methods, what we might get is string of horrendously complicated policies. Just sampling from one of these policies might be tricky. -s

6.2. Function approximation and the need for exploration.

In general, we might not have the information necessary to construct a good class Π_θ, and so the improvement direction must be learned from the data. The function approximation strategy is then to align f with A. This strategy has some powerful advantages, since we can now attempt to jointly increase ω and η. *Actually, I think all the above strategies can be tuned to do that as well -John* First, let us analyze the conditions for improvement in terms of the generalization error.

In practice, the l_2-error of f's approximation to A is a convenient error to minimize. The following theorem gives conditions for improvement based on the ε sub-optimality and the true l_2-error.

This theorem statement feels wrong to me. We can let $\forall s: J(s) = J^*(s)$, and then $E_\nu Af = 0$, breaking the theorem. -John

Theorem 6.1. *Let $\varepsilon > \max_s |J(s) - J^*(s)|$ and $\nu_0 = \min_{s,a} \nu(s,a)$. If*

$$E_\nu(f-A)^2 < \nu_0 \varepsilon^2 (1-\gamma)^2$$

then $E_v Af > 0$.

Proof. Since $E_\nu(f-A)^2 = E_\nu A^2 - E_\nu fA + E_\nu f^2$, then $E_\nu fA \geq E_\nu A^2 - E_\nu(f-A)^2$. Hence, if

$$E_\nu A^2 > E_\nu(f-A)^2 \tag{6.1}$$

, then $E_v Af > 0$. By equation 5.5,

$$E_\nu A^2 \geq \nu_0 \varepsilon^2 (1-\gamma)^2$$

and our result follows. □

Note the ν_0 dependence which suggests that exploration is also necessary for function approximation. If ν_0 is uniform, then a polynomial small error, $\frac{\varepsilon^2(1-\gamma)^2}{n}$, is required to improve ω. Unfortunately, if ν is based on the current policy, then ν_0 could be exponentially small, which implies that an exponentially small generalization error might be required to improve η. Not surprisingly, the exploration problem is also persists for the function approximation case with η.

Let ν and $\tilde{\nu}$ be the distributions implied by μ and μ_e respectively. Again, the problematic case for improving η is if $E_\nu A^2$ is small. However, if we are not in this case we might desire to improve both measures with a reasonable error. Consider

$$E_{\tilde{\nu}}(f-A)^2 + E_\nu(f-A)^2 \leq \tilde{\nu}_0 \varepsilon^2 (1-\gamma)^2$$

so that the error under both error metrics is $\tilde{\nu}_0 \varepsilon^2 (1-\gamma)^2$ small. Note that $E_\nu A^2 > \tilde{\nu}_0 \varepsilon^2 (1-\gamma)^2$ implies η will also improve by equation 6.1. If ν_0 is uniform and the policy is ε sub-optimal, then in addition to improving ω with polynomially small error, η will also improve if $E_\nu A^2$ is not $\frac{\varepsilon^2(1-\gamma)^2}{n}$ small. Thus, in principle, we have lost very little by improving ω.

These results pose the function approximation problem as a supervised learning problem with reasonable requirements. There are two concerns. We need a supervised learning method sophisticated enough to meet these desired goals of a polynomially small generalization error. The other concern is that the representation of the policy can become arbitrarily complicated, since it is the product of all previous f's. In the discussion, we suggest methods to tackle both of these difficulties.

6.3. **What have we gained in the incremental framework?** *Again, the theoretical analysis is not done yet. -s*

OUTLINE: Bounds exist for doing dynamic programming with function approximation, where the algorithm is just to use the greedy update rules specified in DP with approximate value functions. These bounds state how close the final policy is to optimal in terms of the ∞-error. Since we are constructing approximate value functions, the question arises as to why we don't just run DP with these approximate value functions? What have we gained by using incremental improvement? I have not worked it through yet, but I think we gain a factor of $1-\gamma$. In DP with approximation, the closeness of the final policy to optimal is the ∞-error *compounded* $\frac{1}{1-\gamma}$ times. I think in our incremental framework, the final policy is as close to optimal by just the ∞-error, which is the best we could hope for. I haven't finished working it out yet.

7. Discussion

(1) discuss sensible supervised learning methods.
(2) how to represent compact policies?
(3) discuss why the assumption that we have access to an exploration policy is reasonable.

References

8. 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 $\partial_{sa} \equiv \partial/\partial\theta_{sa}$,

$$\partial_{sa}\eta(\pi) = d_\pi(s)\pi(a|s)(Q_\pi(s,a) - J_\pi(s)) = d_\pi(s)\pi(a|s)A_\pi(s,a) \tag{8.1}$$

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

Proof. Let $S^* = \{s|\exists\pi \text{ s.t. } \rho_\pi(s) \neq 0\}$. 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 $\tilde{a} \in \operatorname{argmax}_a Q^{\pi_\infty}(s,a)$, $\partial_{sa}\eta^t$

is positive for sufficiently large t, so $\pi_\infty(a|s) \neq 0$ 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 $\partial_{s\tilde{a}}\eta^t < 0$, 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 8.1 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 $A^\infty(s) \neq 0$. □

9. 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.