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.

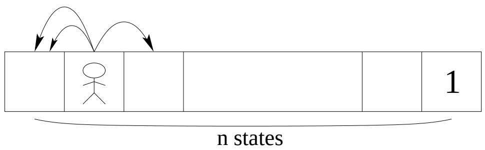

FIGURE 1.1. Example 1: Two actions move agent to the left and one actions moves agent to the right.

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 a Markov decision process 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, we show that determining if any direction (in policy space) will improve the average reward can be arbitrarily difficult.
- In section 6, an alternative performance measure is proposed which is sensitive to performance improvement at unlikely states (unlike the average reward). We show that improvement of this alternative performance measure can never have an arbitrarily small signal to noise ratio.
- In section 7, we discuss how we can improve this alternative performance measure without ever detecting a decrease in the average reward.
- 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 finite Markov decision process (MDP) with rewards. An MDP with rewards is a tuple (S, m_0, A, R, P) where: S is finite set of states, m_0 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:

$$\eta(\pi) = E_{m_0}\{J_\pi(s)\}\,.$$

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 is:

This isn't quite right -John

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

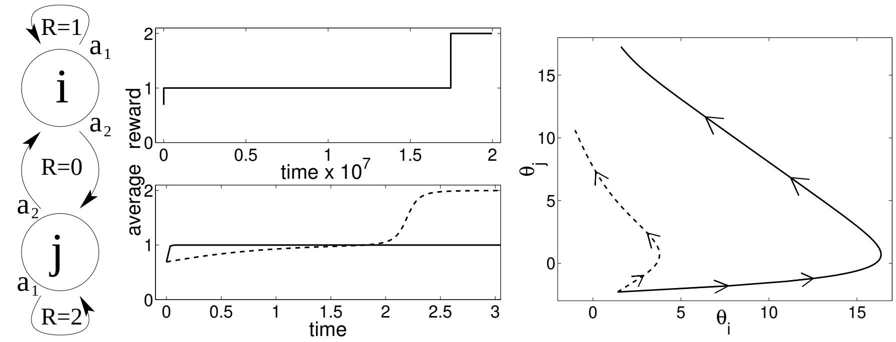

FIGURE 3.1. Example 2: A) See text. B top) The average reward vs. time (on a 10^7 scale) of a policy under standard gradient descent using the sigmoidal policy parameterization ($\pi(1; s, \theta_i) \propto \exp(\theta_i)/(1+\exp(\theta_i))$), with the initial conditions $\pi(i, 1) = .8$ and $\pi(j,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

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:

$$d_{\pi,t}(s) = \Pr_{\pi,\mathrm{MDP}}(s_t = s|s_0) \Pr_{m_0}(s_0)$$

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:

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

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:*

$$\Delta\eta = \alpha < A, f >_{d_\pi \pi} + O(\alpha^2)\,. \tag{4.2}$$

Proof. The gradient of the discounted average reward from a starting state s_0 chosen according to m_0 is:

$$\nabla\eta(\pi) = \nabla E_{s_0} J_\pi(s_0) \tag{4.3}$$

$$= E_{s_0} \nabla E_{\pi(a|s_0)} Q_\pi(s_0, a)$$

$$= E_{s_0}\left[\sum_a Q_\pi(s_0,a)\nabla\pi(a|s_0) + E_{\pi(a|s_0)}\nabla Q_\pi(s_0,a)\right]$$

$$= E_{s_0}\left[\sum_a Q_\pi(s_0,a)\nabla\pi(a|s_0) + E_{\pi(a|s_0)}\nabla\gamma \Pr(s_1|s_0,a)J(s_1)\right]$$

$$= E_{s_0}\left[\sum_a Q_\pi(s_0,a)\nabla\pi(a|s_0) + E_{\pi(a|s_0)}E_{s_1|s_0 a}\gamma\nabla J(s_1)\right]$$

$$= \sum_{t=0}^{\infty}\gamma^t E_{s_t}\left[\sum_a Q_\pi(s_t,a)\nabla\pi(a|s_t)\right]$$

$$= \frac{1}{1-\gamma} E_{s\sim d_\pi(s)}\left[\sum_a Q_\pi(s,a)\nabla\pi(a|s)\right] \tag{4.4}$$

Note: the lack of normalizability shows up here as well with the unavoidable $\frac{1}{1-\gamma}$ term. -John

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:

$$\partial_{sa}\eta(\pi) = \frac{1}{1-\gamma} d_\pi(s)\left[\sum_a Q_\pi(s,a)\pi(a|s)(\delta_{aa'} - \pi(a'|s))\right] \tag{4.5}$$

$$= \frac{1}{1-\gamma} d_\pi(s)\pi(a|s)\left[Q_\pi(s,a) - E_{\pi(a'|s)}Q_\pi(s,a')\right] \tag{4.6}$$

$$= \frac{1}{1-\gamma} d_\pi(s)\pi(a|s)A_\pi(s,a)\,. \tag{4.7}$$

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

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

to first order. $\square$

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:

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

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

$$\frac{\partial\eta}{\partial\theta} = \frac{1}{1-\gamma}E_{s\sim d_\pi(s)}\left[\sum_a \frac{\pi(a|s)}{\pi(a|s)}Q_\pi(s,a)\frac{\partial\pi(a|s)}{\partial\theta}\right]$$
$$= \frac{1}{1-\gamma}E_{s,a\sim d_\pi(s),\pi(a|s)}\left[Q_\pi(s,a)\frac{1}{\pi(a|s)}\frac{\partial\pi(a|s)}{\partial\theta}\right]$$
$$= \frac{1}{1-\gamma}\left\langle\frac{\nabla\pi}{\pi}, Q\right\rangle_{d\pi}\,.$$

The update rule is

$$\pi(a|s) \propto \pi(a|s) + \frac{\partial\pi}{\partial\theta}\Delta\theta$$
$$= \pi(a|s)(1 + \Delta\theta\frac{\partial\log\pi(a|s)}{\partial\theta})$$

and so $f = \frac{\partial \log \pi}{\partial \theta}$. Thus we see that gradient descent is a special case of the previous result in which the direction followed is $\partial \log \pi / \partial \theta$ or $-\partial \log \pi / \partial \theta$ depending on the sign of $< f, Q >_{d\pi}$.

Thus, in order to improve the policy, it is crucial to obtain the correct sign of $< \partial \log \pi / \partial \theta, Q >_{d\pi}$ on average. As stated in the previous section, the variance in obtaining the correct sign can be large.

5. A Different Performance Measure

I think our thoughts might have diverged in our previous section 5&6. You said you failed to see the point in the second part of the previous Thm. 5.1 (the $|\eta - \eta^*|^2 \leq \frac{\langle A,A \rangle_{\tilde{d}\pi}}{m_-^2}$ *part). This second part, which is now in section 5.2, is the hammer for bashing gradient methods. You made an odd comment about polynomial samples needed for* ω*, which seemed to miss the point?*

Also, I slightly changed this section (proofs, thms., etc.) and did not assume the particular choice $f = A$ *when constructing a lower bound. This assumption seemed a bit odd this early on. -Sham*

As discussed in section 3, the agent must have access to a state to improve the policy at that state and the performance measure being used must be sensitive to improvements at those states. 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 m_0), such as

$$\omega(\pi) = \sum_s m_e(s,a) Q_\pi(s).$$

where m_e is some 'exploratory' starting *state-action* distribution. 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. Note that the policy that maximizes ω also simultaneously maximizes the future reward from every state. *Hence, maximizing* ω *also maximizes* η.

The next subsections are outlined as follows:

- in 5.1, we bound the sample size necessary to improve ω
- in 5.2, we argue why using m_e different from m_0 is important
- in 5.3, we discuss the implications of improving ω rather than η

5.1. The Difficulty of Improving ω.

Now let us consider for an arbitrary m_e how difficult is it to improve ω, with the special case that $\omega = \eta$ when $m_e = m_0$. More precisely, 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. Given certain assumptions 8, it is reasonable to suppose that we can calculate a quantity $\hat{Q}(s,a)$ which is an unbiased estimate of the average T-step value of taking action a in state s given policy π. Note that this value is relative to the current value of the policy. Given that all rewards are finite, 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:

$$\tilde{d}_\pi(s) \equiv \frac{1}{T}\sum_{\tau=0}^{T}\sum_{s_0,a_0} m_e(s_0,a_0)\sum_s P(s_\tau = s|s_0,a_0)\,.$$

Under theorem 4.1, ω will improve when the policy π is perturbed toward f if:

$$\left\langle Q, \tilde{f} \right\rangle_\nu > 0$$

, where $\nu(s,a) = \tilde{d}_\pi(s)\pi(a|s)$ which is based on m_e rather than on $m_0(s)$.

Theorem 5.1. *Assume that samples s,a are drawn from $\nu(s,a) \equiv \tilde{d}(s)\pi(a|s)$ and assume that $\left\langle \tilde{f}, \tilde{f} \right\rangle_\nu = 1$. For each sample, we are given an unbiased estimate $\hat{Q}(s,a)$, of $Q(s,a)$ with variance $\sigma^2(s,a)$. Let $\sigma_0 = \max_{s,a}\sigma(s,a)$ and $\sigma_\infty = \max_{s,a}\sigma(s,a)$. Then an unbiased estimate of $< A, f >$ is given by $\hat{Q}(s,a)\tilde{f}(s,a)$ and the variance of this estimate is bounded by:*

$$\sigma_0^2 - \left\langle A, \tilde{f} \right\rangle_\nu^2 \leq \Sigma^2 \leq \sigma_\infty^2 + R_{max}^2\,.$$

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:

$$\Sigma^2 = \left\langle \tilde{f}^2 \hat{Q}^2 \right\rangle_\nu - \mu^2 \tag{5.1}$$

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

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

We can lower bound σ as follows:

$$\begin{aligned}
\sigma^2 &\geq \sum_{s,a}\nu(s,a)\tilde{f}^2\sigma^2(s,a) - \mu^2 \\
&\geq \sigma_0^2\sum_{s,a}\nu(s,a)\tilde{f}^2(s,a) - \mu^2 \\
&= \sigma_0^2 - \mu^2
\end{aligned}$$

where have used the unit length of $\tilde{f}$. We can upper bound σ as follows:

$$\begin{aligned}
\sigma^2 &\leq \sum_{s,a}\nu(s,a)\tilde{f}^2(Q^2(s,a) + \sigma^2(s,a)) \\
&\leq (\sigma_\infty^2 + R_{max}^2)\sum_{s,a}\nu(s,a)\tilde{f}^2(s,a) \\
&= \sigma_\infty^2 + R_{max}^2\,.
\end{aligned}$$

□

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 $\mu = \left\langle A, \tilde{f} \right\rangle_\nu$. The difficulty of this is determined by how much larger the "signal" is than the "noise", *ie* the variance of the estimate of the signal.

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

$$\frac{\Sigma^2}{\mu^2} \geq \frac{\sigma_0^2}{\langle A, A \rangle_\nu} - 1\,.$$

Proof. The $\tilde{f}$ maximizes $\left\langle A, \tilde{f} \right\rangle$ also minimizes $\frac{\Sigma^2}{\mu^2}$ (subject to the normalization constraint on $\tilde{f}$). This this occurs when $\tilde{f} = \frac{A}{\sqrt{\langle A, A \rangle_\nu}}$. Substitution leads to the result. □

Lets consider the special case of $\omega = \eta$ when $m_e = m_0$ and ν is defined through m_0. If all the large advantages are at states that are visited infrequently then $\langle A, A \rangle_\nu$ 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 both of the previous examples. In the two state MDP, at the plateau, state j is visited highly infrequently, and so $\langle A, A \rangle_\nu$ is small, eventhough $\langle A, A \rangle_\nu$ was large to begin with.

Alternatively, 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. Let c

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*

$$\Pr(sign(X) = sign(\left\langle A, \tilde{f} \right\rangle_\nu)) \leq \frac{\sigma_\infty^2 + R_{max}^2}{k \left\langle \tilde{f}, A \right\rangle_\nu^2}$$

Proof. Applying Chebyshev's inequality implies, $\Pr(|X - \mu| > \mu) \leq \frac{\Sigma^2}{k\mu^2}$. The result follows from theorem 5.1. □

I am a bit confused by the following Hoeffding bound. Where did the σ_∞ go? -Sham

We can do better than Chebyshev's inequality using the independence of the $\hat{Q}(s, a)$. In particular, assume that $Q(s, a) \in [0, R_{\max}]$. Then, the Hoeffding bound states:

$$\Pr(|X - \mu| > \mu) \leq e^{-\frac{k\mu^2}{R_{\max}^2}}$$

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

*I disagree with this last argument (from the previous draft). The variance of both policy gradients and our estimates look *equally bad* when $\langle A, A \rangle_{??}$ under* *their respective metric is small (needing $\frac{n^2T^2}{A^2}$ samples). The point of that $|\eta - \eta^*|$ statement is to ask how close are we to the optima when we have a hard time estimating if f improves ω. For η, the answer is we could be very far. For ω, the answer is that we must be close to the optima for a good m_e. The next section discusses this. -Sham*

5.2. **Why is ω better?** The previous section suggested that if $< A, A >_\nu$ 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* m_e) being used. The important question we now address is how difficult is it to determine if f improves a sub-optimal policy. 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.

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

Theorem 5.4. *Let $J^*(s)$ be the optimal reward starting from state s and let $\nu_0 = \min_{s,a} \nu(s,a)$ for some state-action distribution ν. Then*

$$|J(s) - J^*(s)| \leq \sqrt{T\frac{\langle A, A\rangle_\nu}{\nu_0}}. \tag{5.4}$$

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_∞. By induction, we can do this at most for T steps, and so $|J(s) - J^*(s)| < TA_\infty$ (*this is proved in some book)*. The largest possible A_∞ is when $< A, A >_\nu = \nu_0 A_\infty^2$. *I'll construct an example to show this bound is tight in the appendix.* □

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

$$|\omega - \omega^*| \leq \sqrt{T\frac{\langle A, A\rangle_\nu}{\nu_0}} \tag{5.5}$$

where ν is arbitrary (not necessarily based on m).

This bound elucidates the problems with using m_0 as the starting distribution. Let us consider the case in which $\langle A, A\rangle_\nu$ is small, so estimating if f or $-f$ improves the policy will be difficult, by corollary 5.2. For this m_0, the induced state-action distribution can have a small ν_0, so $\langle A, A\rangle_\nu / \nu_0$ could be large (even if $\langle A, A\rangle_\nu$ 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 those important states (making improvement difficult), but the presence of large advantages leaves open the possibility that the policy is significantly sub-optimal.

Instead, let us consider an m_e that does not allow ν_0 to be extremely small. This will force each $J(s)$ to be close to optimal if $\langle A, A\rangle_\nu$ is small. An additional desired property is that the number of samples needed to improve the performance measure must be small. The number of samples must depend on the direction f being followed, and there are clearly high variance choices (those that "alias" states). *However, the following theorem states that there exist f such that local improvement* *can be evaluated with $O(1/\varepsilon)$ samples when the policy could be at most $O(1/\sqrt{\varepsilon})$ away from optimal.* Note that this local improvement of $J(s)$ is with respect to m_e.

Corollary 5.5. *Under the assumptions and sampling model of theorem 5.1, we construct the estimate $X = \frac{1}{k}\sum_{i=1}^{k} \hat{Q}(s,a)\tilde{f}(s,a)$. Then if $\langle A, A\rangle_\nu \leq \varepsilon$ then*

$$|J(s) - J^*(s)| \leq \sqrt{\frac{T\varepsilon}{\nu_0}}$$

else there exist $\tilde{f}$, namely those with $\left\langle \tilde{f}, A\right\rangle_\nu^2 > \varepsilon$, such that

$$\Pr(sign(X) = sign(\left\langle A, \tilde{f}\right\rangle_\nu)) \leq \frac{\sigma_\infty^2 + R_{max}^2}{k\varepsilon}\,.$$

Proof. If $\langle A, A\rangle \leq \varepsilon$, the second condition follows from theorem 5.4. Else if $\langle A, A\rangle_\nu > \varepsilon$, then for $\tilde{f} = A/\sqrt{\langle A, A\rangle_\nu}$, $\left\langle \tilde{f}, A\right\rangle_\nu^2 = \langle A, A\rangle_\nu > \varepsilon$, so such $\tilde{f}$ exist. Using $\left\langle \tilde{f}, A\right\rangle_\nu^2 > \varepsilon$ with corollary 5.3, the second case follows. □

Thus if we use an m_e such that ν_0 is not too small, then each $J(s)$ will be $\sqrt{\frac{T\varepsilon}{\nu_0}}$ near-optimal, when $\langle A, A\rangle_\nu \leq \varepsilon$. By equation 5.5, it follows that η will also be $\sqrt{\frac{T\varepsilon}{\nu_0}}$ near-optimal. If we not guaranteed to be this close to optimal, then there exists a set of directions such that determining if one of these directions improves $J(s)$ (under m_e) needs at most $O(1/\varepsilon)$ samples. *The same cannot be said of policy gradients when the policy is significantly sub-optimal where high variance estimates could occur for a significantly sub-optimal policy.*

5.3. **What about η?** *I weakened this section to essentially state that, in principle, it is possible to 'mostly' improve η when ω improves. We can discuss how in the next section. -Sham*

The previous section suggested that as long as η is not guaranteed to be $\sqrt{\frac{T\varepsilon}{\nu_0}}$-near optimal, then there is the possibility of improving the value functions on average, under m_e, with $O(1/\varepsilon)$ samples. This hinges on finding directions such that $\left\langle \tilde{f}, A\right\rangle_\nu^2 > \varepsilon$ when $\langle A, A\rangle_\nu > \varepsilon$. Clearly, the direction $\tilde{f} \propto A$ and $\tilde{f} \propto -A$ satisfy this criterion. In the next section, we discuss how to find such $\tilde{f}$. Before this, we will address the important issue of what have lost in improving ω instead of η, which is what we wanted to maximize to begin with. *Although, at the end-point of the optimization procedure, we will have found a policy with return close to η^*, an important question is can η decrease along the way until we find 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 starting from m_e and $f(s,a) = -A(s,a)$ on all other states. This will force $\langle f, A\rangle_\nu > 0$ for the ν implied by m_e. However, if the state distribution from m_0 has high probability at states where $f(s,a) = -A(s,a)$ then η can decrease. Thus, although ultimately the policy will be near-optimal, improvement of ω will not always imply improvement of η.

However, there exists $f(s,a)$ for which optimizing $\omega(\pi)$ always agrees with optimizing $\eta(\pi)$. To see this, simply make $f(s,a) = A(s,a)$ everywhere. Then, optimization under any distribution over state-actions will always be in the same direction. Of course, in practice the estimation problem of finding $f(s,a) = A(s,a)$ everywhere is infeasable, which we discuss later.

In general, forcing η to improve at every step is impractical , since determining if any f or $-f$ improves the policy can be arbitrarily difficult for a policy with payoff far from optimal. These difficult cases are when the agent rarely visits states with large advantages, so the reward surface is at a plateau. *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.

Let ν and $\tilde{\nu}$ be the distributions implied by m_0 and m_e respectively and let us consider the case in which we are not guaranteed to be $\sqrt{\frac{T\varepsilon}{\nu_0}}$-close to the optimal policy, *ie* $\langle A, A\rangle_{\tilde{\nu}} \leq \varepsilon$. Corollary 5.5 states that either we can determine if $\tilde{f}$ or $-\tilde{f}$ improves ω with $O(1/\varepsilon)$ samples, for normalized $\tilde{f}$ with $\left\langle \tilde{f}, A\right\rangle_{\tilde{\nu}}^2 > \varepsilon$. As far as η is concerned, let us consider the case in which the agent also visits states with large advantages under m_0, *ie* $\langle A, A\rangle_\nu > \varepsilon$ (*ie* the non-problematic case). Then there exist normalized $\tilde{f}$ such that $\left\langle \tilde{f}, A\right\rangle_{\nu}^2 > \varepsilon$ (for instance $\tilde{f} = A/\sqrt{\langle A, A\rangle_\nu}$) such that we detect improvement with $O(1/\varepsilon)$ samples. Further, there are $\tilde{f}$ that satisfy *both* of these conditions, $\left\langle \tilde{f}, A\right\rangle_{\nu}^2 > \varepsilon$ and $\left\langle \tilde{f}, A\right\rangle_{\tilde{\nu}}^2 > \varepsilon$ (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 m_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. We now discuss how to find these good directions f.

6. The Need for Non-Parametric Function Approximation

Intuitively, when optimizing according to $\omega(\pi)$, we need a family of f which is sufficiently adaptive that optimization on the set of frequently visited states (under $d(s)\pi(a|s)$) is (mostly) orthogonal to optimization under the uniform state distribution given by $\omega(\pi)$. Since the goal state distribution is unknown in advance, this strongly suggests that nonparametric function approximation techniques should be applied. The next theorem essentially states that a reasonable notion of sufficiently adaptive function approximation will work.

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 $\tilde{\Pi}_G = \{\pi_\theta : \pi(a|s) \propto \exp(\theta_{sa})\}$. The exact gradient of the average reward is:

$$\nabla\eta(\pi) = \sum_{s,a} d_\pi(s)\nabla\pi(a|s)Q_\pi(s,a) \tag{7.1}$$

(see [?, ?]). Using the notation $\partial_{sa} \equiv \partial/\partial\theta_{sa}$, we get:

$$\partial_{sa}\eta(\pi) = d_\pi(s)\pi(a|s)(Q_\pi(s,a) - J_\pi(s)) = d_\pi(s)\pi(a|s)A_\pi(s,a) \tag{7.2}$$

Before we address variance issues, we address the important question as to whether or not the the optimal policy will be found by following the gradient. 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 non-trivial 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 $S^* = \{s|\exists\pi \text{ s.t. } \rho_\pi(s) \neq 0\}$. 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 $\tilde{a} \in \text{argmax}_a Q^{\pi_\infty}(s,a)$, $\partial_{sa}\eta^t$ 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 $s_0, a_0, s_1, a_1 \ldots s$, $\rho^t(s) \geq \rho^t(s_0)p^{a_0}_{s_0s_1}\pi_t(a_0|s_0)\ldots p^{a_{\tau-1}}_{s_{\tau-1}s}\pi_t(a_{\tau-1}|s_{\tau-1})$. For every τ-length trajectory with $p^{a_i}_{s_is_{i+1}} \neq 0$ 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 7.2 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 $A^\infty(s) \neq 0$. □

8. How to Monte Carlo

The exact process of how a monte carlo estimate can be formed isn't well spelled out. This is a first attempt. -John

In order to determine which direction is the "right" direction, we need to know how to estimate this right direction. This is a surprisingly subtle problem in the undiscounted average reward case.

The problem of estimating the average reward given a policy and a reinforced MDP is similar to the problem of Monte-Carlo integration of an unknown function with an unknown mixing time. This problem is typically solvable under very general conditions. Let $\hat{p}_s$ be the observed portion of the time the system enters the state s after T steps and p_s be the expected number of times the system enters the state s.

The following condition is a statement that every empirically observed state frequency converges to the exact state frequency at a reasonable rate.

$$\forall s \; \forall \epsilon \; \exists T' \Pr_{\text{T step Markov Process}} (|\hat{p}_s - p_s| \geq \epsilon) \leq e^{-\epsilon^2 \frac{T}{T'}}$$

For a finite state space, this condition implies uniform convergence (over all states simultaneously) at a rate dependent on the size of the state space. This is a very general condition - even finite deterministic systems have a T'. Given a reasonable convergence rate (which is a stronger assumption than the ergodicity assumption) and finite reward values, it is possible to estimate the average reward η_π of the policy to some finite precision in a finite amount of time. Let $\hat{\eta}_\pi$ be the observed average reward.

Intuitively, if we run the Markov process for a time long compared to the convergence time T', we will obtain an estimate $\hat{Q}(s,a)$ which consists of two components, an average reward plus a transient. By subtracting the average reward, we can gain an (approximately unbiased) estimate of the transient.