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.

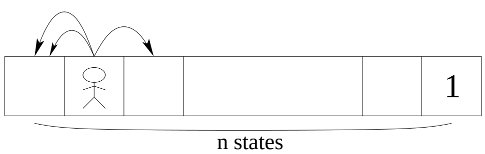

Figure 1.1. Example 1: Two actions move agent to the left and one actions moves agent to the right.

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 T-step reward from s_0, which can be seen as an approximation scheme to maximizing the average reward. We define the state-action value as

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

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:

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

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 $O(2^n)$ 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 $O(2^{-n})$. This, implies that the expected number of steps between visits to the goal state is $O(2^n)$. 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 commonly used Gibbs table-lookup distribution, $\tilde{\Pi}_G = \{\pi_\theta : \pi(a|s) \propto \exp(\theta_{sa})\}$. Using the notation $\partial_{sa} \equiv \partial/\partial\theta_{sa}$, it is straightforward to show that the gradient of the average reward is:

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

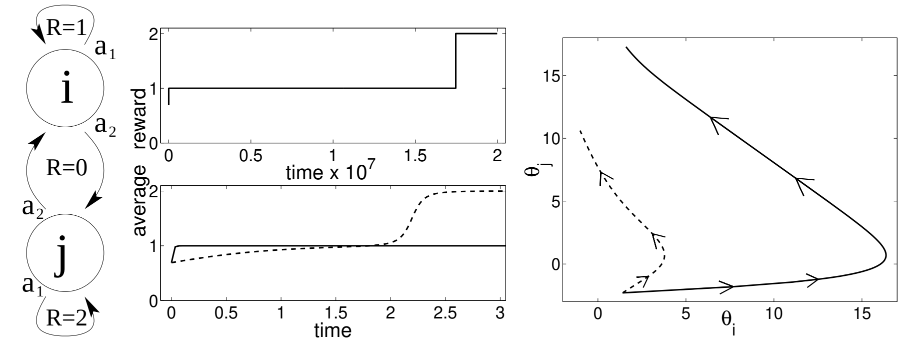

FIGURE 3.1. Example 2: A) See text. B top) The average reward vs. time (on a 10^7 scale) of a policy under standard gradient descent using the sigmoidal policy parameterization ($\pi(1; s, \theta_i) \propto \exp(\theta_i)/(1+\exp(\theta_i))$), with the initial conditions $\pi(i,1) = .8$ and $\pi(j,1)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

Gradient methods guarantee improvement due to the incremental nature of the algorithm although no *rate* of improvement guarantees are made. Dynamic programming using function approximation has only 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. Under the starting state distribution $m_0(s)$, the T-step future distribution of states is:

$$d_\pi(s) \equiv \frac{1}{T}\sum_{\tau=0}^{T}\sum_{s_0} m(s_0)\sum_{s} P(s_\tau = s|s_0)$$

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

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

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

Theorem 4.1. *Assume the definition given above for $d_\pi(s)$. Let us consider the update rule $\pi(a|s) \leftarrow \pi(a|s)(1+\alpha f(s,a))$ for some vector f and some small α. Let A be the true advantages. Then the performance improvement of η is:*

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

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

$$\nabla\eta(\pi) = \sum_{s,a} d_\pi(s)\nabla\pi(a|s)Q_\pi(s,a). \tag{4.3}$$

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

$$\begin{aligned}
\partial_{sa}\eta(\pi) &= \sum_a d(s)\pi(a|s)(\delta_{aa'} - \pi(a'|s))Q(s,a) && (4.4)\\
&= d_\pi(s)\pi(a|s)(Q_\pi(s,a) - < Q_\pi(s,a') >_{\pi(a'|s)}) && (4.5)\\
&= d_\pi(s)\pi(a|s)A_\pi(s,a)\,. && (4.6)
\end{aligned}$$

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 $\Delta\eta = \sum_{sa}\partial_{sa}\eta\Delta\theta_{sa} = \sum_{sa} d(s)\pi(a|s)A(s,a)\Delta\theta_{sa}$ (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

$$\Delta\eta = \sum_{sa} d(s)\pi(a|s)A(s,a)(\alpha f_{sa}) = < A, f >_{d\pi}$$

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:

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

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 update rule under a gradient method is

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

and so

$$f = \frac{\partial \log \pi}{\partial \theta}$$

It follows that if $\langle \partial \log \pi / \partial \theta, A \rangle_{d\pi} > 0$ then a small positive $\Delta\theta$ will improve the policy, otherwise a negative $\Delta\theta$ will improve the policy.

Under gradient descent, $\Delta\theta$ will have the same sign as $\partial\eta/\partial\theta$. The gradient (equation 4.3) can be written in the form:

$$\partial\eta/\partial\theta = < \partial \log \pi/\partial\theta, Q >_{d\pi} \quad .$$

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 $< \partial \log \pi/\partial\theta, Q >_{d\pi}$.

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. The next section discusses the variance issue in determining the sign of this quantity.

5. Fixing Gradient Techniques

TODO: Merge sections 5&6. Merge upper and lower bounds on noise to signal. Split off statements about η.

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 reinforced MDP. Given certain assumptions 9, 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:

$$\forall s \; E_{\pi(a|s)} \tilde{f}(s,a) = 0$$

The theorem will use the notion of "signal" and "noise" where the signal is given by:

$$\mu = E_{d(s)\pi(a|s)} \tilde{f}(s,a) Q(s,a) = E_{d(s)\pi(a|s)} \tilde{f}(s,a) A(s,a)$$

and the noise (variance) is given by:

$$\Sigma^2 = E_{d(s)\pi(a|s)}\tilde{f}^2(s,a)Q^2(s,a) - \mu^2\,.$$

Theorem 5.1. *Assume that samples s,a are drawn from $d(s)\pi(a|s)$. For each sample, we are given an unbiased estimate $\hat{Q}(s,a)$, of $Q(s,a)$ with variance $\sigma^2(s,a)$. Let $\sigma_0 = argmin_{s,a}\sigma(s,a)$. Then for all vectors $f(s,a)$ the noise to signal ratio is bounded by:*

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

Further, if η^* is the optimal possible average reward and $m_- = \min_{s,a} d_\pi(s)\pi(a|s)$, then

$$|\eta - \eta^*|^2 \leq \frac{\langle A,A\rangle_{d\pi}}{m_-^2}\,.$$

Intuitively, this result states that if the advantages are small on average, under $d\pi$, then the gradient will be difficult to estimate. For this to occur, any large advantages must be at states that are visited infrequently. 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 $\langle A,A\rangle_{d\pi}$ is small (eventhough $\langle A,A\rangle_{d\pi}$ was large to begin with). *Unfortunately, if large advantages are present (even at unlikely states), the agent may gain a large performance increase.*

Proof. Since the $\hat{Q}$ are unbiased, the mean μ is:

$$\begin{aligned}\mu &= \sum_{s,a} d(s)\pi(a|s)\int dP(\hat{Q}(s,a))\hat{Q}(s,a)\tilde{A}(s,a)\\ &= \left\langle \tilde{f}, Q\right\rangle_{d\pi} = \left\langle \tilde{f}, A\right\rangle_{d\pi}\end{aligned}$$

and the variance is:

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

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

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

giving us:

$$\frac{\Sigma^2}{\mu^2} = \frac{\left\langle \tilde{f}^2(Q^2+\sigma^2)\right\rangle_{d\pi}}{\left\langle \tilde{f}, A\right\rangle_{d\pi}^2} - 1$$

Technically, $\tilde{f}$ is constrained such that $\sum_a \tilde{f}(s,a)\pi(a|s) = 0$ (or equivalently $\tilde{f}(s,a) \equiv f(s,a) - \sum_{a'}\pi(a'|s)f(s,a')$). The goal is to minimize $\frac{\Sigma^2}{\mu^2}$ subject to this constraint. However, let's just minimize this quantity unconstrained since that gives a bound. Thus:

$$\frac{\partial(\sum^2/\mu^2)}{\partial \tilde{f}_{sa}} = \frac{2d(s)\pi(a|s)\tilde{f}(s,a)(Q^2(s,a)+\sigma^2(s,a))}{\mu^2} - \frac{2\left\langle \tilde{f}^2(Q^2+\sigma^2)\right\rangle_{d\pi}\frac{\partial\mu}{\partial \tilde{f}_{sa}}}{\mu^3}$$

$$= \frac{2d(s)\pi(a|s)\tilde{f}(s,a)(Q^2(s,a)+\sigma^2(s,a))}{\mu^2} - \frac{2d(s)\pi(a|s)A(s,a)\left\langle \tilde{f}^2(Q^2+\sigma^2)\right\rangle_{d\pi}}{\mu^3}$$

We can verify that $\tilde{f}^*(s,a) = \frac{A(s,a)}{Q^2(s,a))+\sigma^2(s,a)}$ is an extrema.

By substitution, $\mu(\tilde{f}^*) = \left\langle \frac{A^2}{Q^2+\sigma^2}\right\rangle_{d\pi}$. Also, since $\sum_{s,a} d(s)\pi(a|s)(\tilde{f}^*)^2(Q^2(s,a))+\sigma^2(s,a)) = \mu(\tilde{f}^*)$, it follows that :

$$\Sigma(\tilde{f}^*)^2 = \mu(\tilde{f}^*) - \mu(\tilde{f}^*)^2\,. \tag{5.4}$$

Using $\Sigma^2/\mu^2 \geq 1/\mu(\tilde{f}^*) - 1$, it follows that

$$\frac{\Sigma^2}{\mu^2} \geq \frac{1}{\left\langle \frac{A^2}{Q^2+\sigma^2}\right\rangle_{d\pi}} - 1\,.$$

Our result follows from using the definition of σ_0.

The second part of the proof is easy.

I'm not sure what the point of the second part is, but maybe seperate and fill in -John □

We view this variance trap as problem of exploration. 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.

6. A Different Performance Measure

We desire a performance measure that can never have an arbitrarily high signal to noise ratio when there is a potential for large performance increase (*ie* when there are large advantages at some states). This alternative performance measure cannot down weight advantages at unlikely states or unlikely actions. The performance measure we choose will *require* access to an exploration policy which can achieve a "uniform" distribution over states. This requirement is both nontrivial and feasible.

A natural candidate for a performance measure is to weight the improvement from all states and all actions more uniformly, 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, η is also maximized).

Now let us analyze the improvement conditions for this new performance measure and its signal to noise properties. The analogous discounted future distribution of states is:

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

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

$$\langle A, f\rangle_{\tilde{d}\pi} > 0$$

, where, crucially, the metric uses $\tilde{d}_\pi(s)$ instead of $d_\pi(s)$.

The following theorem shows that there exists a direction in which improvement of ω can always be detected (with a policy independent bound), unless the policy has performance close to the globally optimal policy.

Theorem 6.1. *Assume that samples s, a are sampled from $\tilde{d}(s)\pi(a|s)$ and that the unbiased Monte Carlo estimates $\hat{Q}(s,a)$ of $Q(s,a)$ have variance $\sigma^2(s,a)$. Let $\sigma_\infty = \max_{s,a}\sigma(s,a)$. If the improvement direction f is the true advantages A, then the noise to signal ratio is bounded by:*

$$\Sigma^2/\mu^2 \le \frac{\sigma_\infty^2}{\langle A, A\rangle_{\tilde{d}\pi}} \,.$$

Further, the if η^* is the optimal possible average reward and $m_- = \min_{s,a} m_e(s,a)$, then

$$|\eta - \eta^*|^2 \le \frac{\langle A, A\rangle_{\tilde{d}\pi}}{m_-^2} \,.$$

Note that for the uniform distribution $m_- = 1/(\#\text{ of states})$, whereas in the theorem 5.1, the related term could be arbitrarily smaller, depending on the policy. Similar to the previous result, the noise to signal ratio could be high if $\langle A, A\rangle_{\tilde{d}\pi}$ is small. *However, small $\langle A, A\rangle_{\tilde{d}\pi}$ forces the policy to have close to optimal performance, unlike the previous situation for small $\langle A, A\rangle_{\tilde{d}\pi}$.*

Proof. Using $f = A$, the mean is

$$\mu = \sum_{s,a} \tilde{d}(s)\pi(a|s)A^2(s,a)$$

and using equation 5.3, the variance of an estimate can be bounded by

$$\begin{aligned}
\sigma^2 &\le \sum_{s,a} \tilde{d}(s)\pi(a|s)\sigma^2(s,a)A^2(s,a) && (6.1)\\
&\le \sigma_\infty^2 \sum_{s,a} \tilde{d}(s)\pi(a|s)A^2(s,a) && (6.2)\\
&= \sigma_\infty^2 \langle A, A\rangle_{\tilde{d}\pi} \,. && (6.3)
\end{aligned}$$

The first result follows using $\sigma^2/\mu^2 \le \sigma_\infty^2/\mu^2$.

outline: The second part follows by bounding A_∞ (the largest advantage). Then it is straightforward to show if the maximum advantage is A_∞ then the optimal payoff can be no greater then $A_\infty T$ over the current policy. □

What does this imply about the number of samples $\hat{Q}(s,a)$? Consequently, the variance will decrease like $\frac{1}{k}$ where k is the number of samples. Let $X = \frac{1}{k}\sum_{i=1}^{k} \hat{Q}(s,a)\tilde{f}(s,a)$. Applying Chebyshev's inequality implies:

$$\Pr(|X-\mu| > \mu) \le \frac{\sigma_\infty^2}{k\,\langle A, A\rangle_{\tilde{d}\pi}}$$

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:

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

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 $k = \frac{n^2T^2}{A^2}$ 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.

7. The Need for Non-Parametric Function Approximation

For an arbitrary f, optimizing $\omega(\pi)$ can have disastrous effects on the average reward, η_π. To see this consider a one parameter function $f(s,a)$ for which $f(s,a) = A(s,a)$ on all states commonly visited under $d(s)\pi(a|s)$ and $f(s,a) = -A(s,a)$ on all other states. The advantages can be essentially independent which implies that given sufficient "other states", optimization according to $\omega(\pi)$ will be directly opposed to opimization of η_π. Therefore, optimizing according to $\omega(\pi)$ for arbitrary f is dangerous.

There exists $f(s,a)$ for which optimizing $\omega(\pi)$ always agrees with optimizing η_π. 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 we will not use $f(s,a) = A(s,a)$.

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.

Let us first consider the signal to noise under a general state-action distribution $\nu(s,a)$. The previous condition only considered the case where $f = A$. Now we consider the case when f is ε close to $\tilde{A}$.

This theorem sort of bothers me, but I'm not certain why. Maybe it's because there are more general families of f which will work. Why does variance go to ∞ when ε goes to 0? -John

Theorem 7.1. *Assume that samples s, a are sampled from $\tilde{d}(s)\pi(a|s)$ and that the unbiased Monte Carlo estimates of $Q(s,a)$ have variance $\sigma^2(s,a)$. Let $\sigma_\infty = \max_{s,a}\sigma(s,a)$. If $\langle f-A, f-A\rangle_\nu < \varepsilon^2$ and $\langle A, A\rangle_\nu > 2\varepsilon^2$, then:*

$$\frac{\Sigma^2}{\mu^2} \leq \frac{2\sigma_\infty^2}{\varepsilon^2}\,.$$

Proof. Using equation 6.3, we have:

$$\frac{\Sigma^2}{\mu^2} \leq \frac{\sigma_\infty^2 \langle A, A\rangle_\nu}{\langle A, f\rangle_\nu}\,.$$

Since $\langle f-A, f-A\rangle_\nu < \varepsilon^2$ and $\langle A, A\rangle_\nu > 2\varepsilon^2$, then $\langle A, f\rangle_\nu > \varepsilon^2$. The result follows from combining inequalities. □

Intuitively, as long as $\langle A, A\rangle_\nu$ is large, then any f that is ε-close to A will have a $1/\varepsilon^2$ bound on its noise to signal ratio.

Now, let us consider the consequences for improvement under both performance measures. Consider f satisfying:

$$\langle f - A, f - A\rangle_{\tilde{d\pi}} + \langle f - A, f - A\rangle_{d\pi} \leq \varepsilon^2 \tag{7.1}$$

, which is the set of vectors that is ε-close to A under both $\tilde{d\pi}$ and $d\pi$. For these vectors, we will either be able to detect improvement of ω or the policy has near optimal performance. Further, we will either be able to detect an increase of η or we are in the situation where the signal to noise ratio is small (and so improving η was hard anyways). More precisely, for the ε-close f,:

(1) If $|A|_{\tilde{d\pi}} > 2\varepsilon$, then f will improve ω and detection of this improvement is feasible ($\Sigma/\mu \leq \frac{\sigma_\infty}{\varepsilon}$). Else if $|A|_{\tilde{d\pi}} \leq 2\varepsilon$, then although detection of improvement is not feasible, the policy is close to optimal ($|\eta - \eta^*| \leq \frac{2\varepsilon}{(1-\gamma)m_-}$).
(2) If $|A|_{d\pi} > 2\varepsilon$, then f will improve η and detection of this improvement is feasible ($\Sigma/\mu \leq \frac{\sigma_\infty}{\varepsilon}$). Else if $|A|_{d\pi} \leq 2\varepsilon$, then detection could be difficult (since $\Sigma/\mu \leq \frac{\sigma_0}{2\varepsilon} - 1$ by theorem 5.1).

The big picture is that we have lost very little by improving ω. The only case where η might decrease when ω improves is when $|A|_{d\pi}$ is small. However, by theorem 5.1, this was the case where it is difficult to improve η anyways due to the high variance, so improvement was an unreasonable expectation. The flip side is that we will not be able to detect a decrease of η , for small changes to the policy (in essence, this is the 'plateau' case). However, what we have gained is that we can always improve ω until the policy is close to optimal (and η along the way, when feasible).

Note that for a fixed architecture, it will not necessarily be possible to satisfy equation 7.1. Thus, we need to consider function approximation schemes that are are rich enough to satisfy this quantity. Intuitively, we hope that policy improvement at unlikely states does not clash with improvement at likely states (when feasible, as discussed above). Further, we seek to eventually find the optimal policy. *This leads into non-parametric schemes!*

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

$$\nabla\eta(\pi) = \sum_{s,a} d_\pi(s)\nabla\pi(a|s)Q_\pi(s,a) \tag{8.1}$$

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

$$\partial_{sa}\eta(\pi) = d_\pi(s)\pi(a|s)(Q_\pi(s,a) - J_\pi(s)) = d_\pi(s)\pi(a|s)A_\pi(s,a) \tag{8.2}$$

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 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 $\tilde{a} \notin \operatorname{argmax}_a Q^{\pi_\infty}(s,a)$ 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 8.2 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 $A^\infty(s) \neq 0$. □

9. 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.

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

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.

10. Have we thrown the baby out with the bath water?

aside: The gist of this section is that following ω does not hurt us with respect to η, except when we are at a 'plateau'. For this case, we are screwed anyways with regards to η, so no big deal.

aside: It should be easy to come up with a case (as specified above in point 2) where η will decrease.