Incremental Reinforcement Learning

19th September 2001

Abstract

Gradient methods fundamentally intertwine exploration and exploitation. We present a simple example showing how samples of the gradient can become arbitrarily noisy. To flesh out the problems, we first consider the general conditions for improvement when incrementally updating a policy. We show that determining if any direction (in policy space) will improve the performance of the current policy can be arbitrarily difficult (*ie* an upper bound on the signal to noise ratio can be shown to be arbitrarily small). We argue that this is the result of a lack of exploration (or, equivalently, an inability to judge if the policy improves at states that have low stationary probability). To overcome these difficulties, an alternative performance measure is suggested which does not suffer from this variance trap (since it can judge improvement of the policy at these unlikely state). Further, we argue that maximizing this alternative performance measure can be done in such a way that we can never detect a decrease in the average reward (where the detection criterion is determined by the noise).

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 [9]. For example, consider the simple MDP shown in Figure 1A (adapted from [8]). Under a policy that gives equal probability of choosing any action, 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

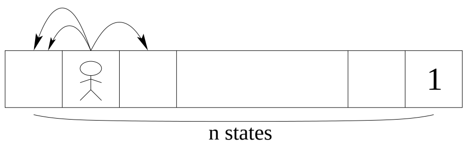

Figure 1: Example 1: Two actions move agent to the left and one actions moves agent to the right.

state. Importance sampling methods have been formulated [4, 5] though it is unreasonable to expect that they can salvage the situation — when there is an exponentially low probability event with a significant contribution to the expectation of interest.

In contrast, 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 backpropagate this information to those states that are visited often (see [2, 8]).

An additional problem is that eventhough the average reward is what we ultimately seek to maximize, this performance measure has the fundamental problem that it is insensitive to performance improvement at highly unlikely states. However, improvement at these states could lead to drastic policy improvement (as in example 1). We formalize this by showing that it can be arbitrarily difficult to improve a policy which has significantly sub-optimal performance. What is meant by difficult is that estimating if a direction (in policy space) will improve the policy could have an arbitrarily small signal to noise ratio.

Again, in contrast, dynamic programming seeks to improve the policy at all states. Thus, unlikely states will exhibit policy improvement eventhough this improvement might have little impact on the average reward.

In section 2, the problems with gradient methods are discussed in more detail (though we present a result showing the *prima facie* picture seems promising). The summary is that finding the *direction* of the gradient when the *magnitude* is small can have high variance. The crucial point is that the magnitude of the gradient can be small at plateaus far away from the optima (*ie* eventhough being close to the optima implies small gradients, small gradients do not imply being close to optima). In section 3, we discuss general incremental policy improvement (and consider both gradient methods and function approximation). In section 4, we formalize the problems of only trying to maximize the average reward by showing that determining if any direction (in policy space) will improve the average reward can be arbitrarily difficult. Again, what is meant by difficult is a low signal to noise ratio. In section 5, we suggest maximizing an alternative performance measure that 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 6, we discuss how we can improve this alternative performance measure without ever detecting a decrease in the average reward (which is what we wanted to maximize in the first place).

2 Following the Gradient

Ultimately, we care about maximizing the average reward, so this seems like the natural performance measure to use. However, before a global optima is reached there can be regions in which it is arbitrarily difficult to improve the average reward.

First, let us state the standard formulation of a finite Markov decision process (MDP). An MDP is a tuple (S, s_0, A, R, P) where: S is finite set of states, s_0 is a start state, 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 $\pi(s,a)$, 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 distribution d^π. Under this assumption, the *average reward* is $\eta(\pi) \equiv \sum_{s,a} d^\pi(s)\pi(s,a)R(s,a)$, the state-action value is $Q^\pi(s,a) \equiv E_\pi\{\sum_{t=0}^\infty R(s_t,a_t) - \eta^\pi | s_0 = s, a_0 = a\}$ and the value function is $J^\pi(s) \equiv E_{\pi(s,a')}\{Q^\pi(s,a')\}$, where and s_t and a_t are the state and action at time t. For lack of space, we only consider the more difficult case where the goal of the agent is to find a policy that maximizes the average reward over policies in $\tilde{\Pi} \subset \Pi$ (where Π is the set of all policies).

In this section, we trivially restrict the class of policies by excluding only deterministic policies and consider only the Gibbs table-lookup distribution, $\tilde{\Pi}_G = \{\pi_\theta : \pi(s,a) \propto \exp(\theta_{sa})\}$. The exact gradient of the average reward is:

$$\nabla\eta(\pi) = \sum_{s,a} d^\pi(s)\nabla\pi(s,a)Q^\pi(s,a) \tag{1}$$

(see [7, 3]). Using the notation $\partial_{sa} \equiv \partial/\partial\theta_{sa}$, it is straightforward to show :

$$\partial_{sa}\eta(\pi) = d^\pi(s)\pi(s,a)(Q^\pi(s,a) - J^\pi(s)) = d^\pi(s)\pi(s,a)A^\pi(s,a) \tag{2}$$

where we define the advantage to be

$$A^\pi(s,a) \equiv Q^\pi(s,a) - J^\pi(s)\,.$$

Before we address variance issues, we address the natural 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, which suggests the dismal possibility that the gradient descent could lead to any deterministic policy. It is straight forward to show that if $d(s)$ is nonzero for all states then the gradient tends to improve the probability actions with positive advantages, and thus the optimal policy will be reached. 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 $\theta_t = \theta_0 + \int_0^t \nabla\eta^{\pi_t} dt$. ***The proof is a bit hairy at the moment (but correct), so don't bother wading through it too deeply. For some reason I used $\rho(s)$ as the stationary distribution instead of $d(s)$. However, the result was quite interesting to me and non-obvious.***

Theorem 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 $\pi_t(s,a) \to 0$ 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 $\pi_\infty(s,a) \neq 0$ 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 $\pi_\infty(s,\tilde{a}) = 0$, 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 $\rho^t \to 0$, $\pi_t(s,a) \to 0$, or $A^t(s,a) \to 0$. We now show that even if if $\rho^\infty \neq 0$, then $\pi_\infty(s,a) = 0$ 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 $s_0, a_0, s_1, a_1 \ldots s$, $\rho^t(s) \geq \rho^t(s_0)p^{a_0}_{s_0 s_1}\pi_t(s_0,a_0)\ldots p^{a_{\tau-1}}_{s_{\tau-1}s}\pi_t(s_{\tau-1},a_{\tau-1})$. 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 $\pi_\infty(s_i,a_i) = 0$, 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 $\forall t\, \rho^t(s) \geq \tilde{\alpha}\rho^t(\tilde{s})\pi_t(\tilde{s},\tilde{a})$, $\rho^\infty(\tilde{s}) \neq 0$, and $\pi_\infty(\tilde{s},\tilde{a}) = 0$ (using a trajectory such that $\pi_\infty(s_i,a_i) = 0$ 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 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 $A^\infty(s) \neq 0$. **q.e.d.**

This suggests that using a stochastic gradient algorithm with the usual caveats is *prima facie* sufficient to find the globally optimal policy. Unfortunately, the variance of a gradient estimate can become arbitrarily large. This is due to the fact that, to estimate $\partial_{sa}\eta$, the state s must sampled under the stationary distribution (which could have a probability that is exponentially small in the size of the state space). Consider the 2-state MDP shown in Figure 1B, increasing the chance of a self-loop at i decreases the stationary probability of j, which hinders the learning to j. . Under an initial policy that has the stationary distribution $\rho(i) = .8$ and $\rho(j) = .2$, learning to state i squelches the learning to state j and this leads to an an extremely flat plateau of improvement at 1

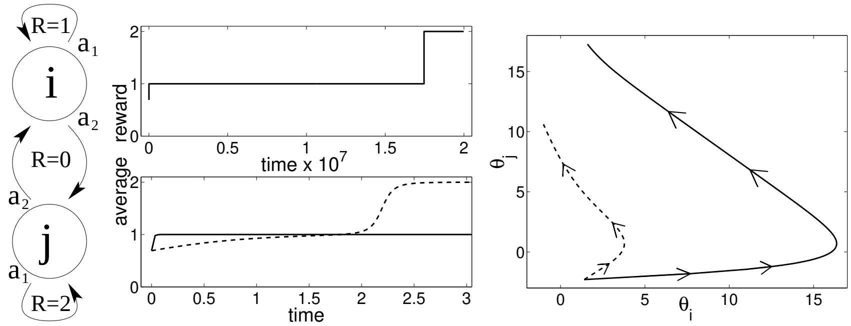

Figure 2: Example 2: A) See text. B top) The average reward vs. time (on a 10^7 scale) of a policy under standard gradient descent using the sigmoidal policy parameterization ($\pi(1; s, \theta_i) \propto \exp(\theta_i)/(1+\exp(\theta_i))$), with the initial conditions $\pi(i, 1) = .8$ and $\pi(j, 1)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

average reward shown in Figure 1C (see figure text). Figure 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 1E shows where this occurs and gives a clearer picture (see figure text).

Let us now examine the variance in estimating the gradient more thoroughly. First, let us state a bound from [1] on the variance of a gradient estimate based on single Monte Carlo run of N steps:

$$Pr(|\Delta - \Delta_{\text{MC}}|_\infty \geq \varepsilon) \geq \frac{Nd}{2} \exp(-s/\tau) + 2sd \exp(-\varepsilon^2 N \frac{c}{s}) \tag{3}$$

where Δ and Δ_{MC} is the true and sampled estimate of the gradient, s is any number less than N, and c is a constant. The norm $|\,|_\infty$ is the largest difference in gradient over all the parameters. *The important point is that the direction needs to be correct (within* 90 *degrees) for improvement regardless of the value of* ε. *The bound shows that we can easily force correctness to within* ε, *but this will not necessarily get the direction correct if the magnitude of the gradient is small. This example shows that the magnitude of the gradient can be small far away from the optima.*

aside: check out the paper. big guns. my simple example and the previous paragraph show no aim. direction is what is important to get the gradient correct.

Reconsidering the MDP in example 1, we see that although it suffers from the initial sampling problem, it is simple in comparison to our 2-state MDP in the sense that the gradient is always improving the probability of the optimal action, unlike in in the 2 state problem where the gradient initially reinforces a sub-optimal action at state i. In other words, one step of policy improvement from any policy returns an optimal policy for the MDP in example 1. However, the placement of small reward at the left-most state would complicate matters even further, since the agent would be encouraged to explore even less. The 2-state MDP demonstrates that even if the initial policy visits all states with reasonable probabilities this does not prevent the gradient from entering a region where the estimates have high variance.

3 Incremental Policy Improvement

Gradient methods guarantee improvement due to the incremental nature of the algorithm. Dynamic programming using function approximation has only weak theoretical guarantees which stem from the fact that drastic changes to the policy could occur. It seems that feasible performance guarantees require the policy to be changed incrementally (by following a gradient or via value function approximations). In the first subsection, general conditions under which an incremental change to the policy will improve its performance are stated, and the later sub-sections analyze how to satisfy these conditions.

3.1 General Conditions for Improvement

This subsection (only 3.1) is pretty much an excerpt from the other paper. However, note the slightly different notation. f is general vector, and $\tilde{A}$ is defined through f (in the next subsection)

Let's now consider the incremental update rule:

$$\pi_s(a) \leftarrow \pi_s(a)(1 + \alpha f(s,a)) \tag{4}$$

where f is some vector. Note that this update rule is to first order equal to the update rule $\pi_s(a) \leftarrow \pi_s(a)\exp(\alpha f(s,a))$, which keeps the probabilities positive. Also, 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.

First, let us define the inner product for vectors in state-action space to be:

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

where $\rho(s)\pi(s,a)$ is state-action distribution. This leads to following result:

Theorem Let us consider the update rule specified in equation 4 for some f and some small α. Let A be the true advantages. Then the performance improvement is:

$$\Delta\eta = \alpha < A, f >_{\nu} + O(\alpha^2) \tag{5}$$

where $\nu(s,a) = d^{\pi}(s)\pi(s,a)$.

Thus, to obtain improvement all we need is the much weaker result that $\tilde{f}$ must have a positive inner product with A (as opposed to strong and intractable conditions with the l_∞ norm).

Proof First, let us consider the table-lookup Gibbs distribution, $\tilde{\Pi}_G = \{\pi_\theta : \pi(s,a) \propto \exp(\theta_{sa})\}$. The exact gradient of the average reward (for any parameterized policy) is:

$$\nabla \eta^\pi = \sum_{s,a} \rho^\pi(s) \nabla \pi(s,a) Q^\pi(s,a). \tag{6}$$

Using the notation $\partial_{sa} \equiv \partial / \partial \theta_{sa}$, it is straightforward to show for $\tilde{\Pi}_G$, since $\partial_{sa} \pi_{s'a'} = \delta_{ss'} \pi(s,a)(\delta_{aa'} - \pi(s,a'))$:

$$\begin{aligned} \partial_{sa} \eta^\pi &= \sum_a \rho(s) \pi(s,a)(\delta_{aa'} - \pi(s,a')) Q(s,a) & (7) \\ &= \rho^\pi(s) \pi(s,a)(Q^\pi(s,a) - < \tilde{Q}(s,a') >_{\pi(s,a')}) & (8) \\ &= \rho^\pi(s) \pi(s,a) A^\pi(s,a) & (9) \end{aligned}$$

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

$$\pi_s(a, \theta_{t+1}) \leftarrow \pi_s(a; \theta_t) \exp(\Delta \theta_{sa})$$

, the resulting change to the performance is $\Delta \eta = \sum_{sa} \partial_{sa} \eta \Delta \theta_{sa} = \sum_{sa} \rho(s) \pi_s(a) A(s,a) \Delta \theta_{sa}$ (for small $\Delta \theta$).

The update rule for the table-lookup Gibbs is exactly the same as that used in equation 4, so we evaluate the change to the performance

$$\Delta \eta = \sum_{sa} \rho(s) \pi_s(a) A(s,a)(\alpha f_{sa}) =< A, f >_\nu$$

. **q.e.d.**

3.2 How do we gain access to the advantages?

How do we estimate $< f, A >_{d\pi}$, since the advantages can't be obtained from Monte-Carlo estimates? The answer is to use the state-action values, which can be estimated from Monte Carlo samples. The advantages are different from these state-action values by a state-dependent constant, $A(s,a) = Q(s,a) - J(s)$.

Let us define the approximate advantages as:

$$\tilde{A}_{sa} \equiv f_{sa} - \sum_{a'} \pi_{sa'} f_{sa'} \tag{10}$$

, so that $\tilde{A}$ is f with the baseline removed.

Theorem Let $\tilde{A}_{sa} \equiv f_{sa} - \sum_{a'} \pi_{sa'} f_{sa'}$. Then:

$$< f, A >_\nu =< \tilde{A}, Q >_\nu$$

where $\nu(s,a) = d(s)\pi(s,a)$.

proof: This follows from:

$$\begin{aligned}
< \tilde{A}, Q >_\nu &= \sum_{s,a} d(s)\pi(s,a)\tilde{A}(s,a)(A(s,a) + J(s)) \\
&= \sum_{s.a} d(s)\pi(s,a)\tilde{A}(s,a)A(s,a) \\
&= \sum_{s,a} d(s)\pi(s,a)(f(s,a) - \sum_{a'} \pi(s,a')f(s,a'))A(s,a) \\
&= \sum_{s,a} d(s)\pi(s,a)f(s,a)A(s,a) \\
&= < f, A >_\nu
\end{aligned}$$

where we have used the fact that $\sum_a \pi(s,a)\tilde{A}(s,a) = 0$ and $\sum_a \pi(s,a)A(s,a) = 0$. **q.e.d.**

3.3 Gradient Methods

For simplicity, let us consider the class of policies parameterized by one parameter, θ. The update rule under a gradient method is

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

and so

$$f = \partial \log \pi / \partial \theta$$

(also $\tilde{A} = \partial \log \pi / \partial \theta$ since probabilities are normalized). It follows that if $< \partial \log \pi / \partial \theta, Q >_\nu > 0$ then a small positive $\Delta\theta$ will improve the policy, else 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 can be written in the form:

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

, which is the importance sampling version of equation 6. 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 $< \partial \log \pi / \partial \theta, Q >_{d\pi}$.

The parameterization chooses a direction, $f = \partial \log \pi / \partial \theta$, and the gradient tells us to move in f or $-f$ direction (for many parameters, the parameterization chooses a subspace). Thus, in order to improve the policy, it is crucial to obtain the correct sign of $< \partial \log \pi / \partial \theta, Q >_{d\pi}$ (not necessarily related to ε in bound 3) . The next section discusses the variance issue in determining the sign of this quantity.

3.4 Function Approximation

An alternative approach is to use function approximation to find some f or $\tilde{A}$ to improve the policy with. Two important question are what should we fit and how should we do it? As far as improvement is concerned, if each $\tilde{A}(s,a)$ has the same sign as $A(s,a)$ then improvement will be guaranteed. In this case, an infinite α will provide a policy with greater performance than the initial policy (though a line search would not necessarily end up at this point, since monotonicity is not guaranteed even if the advantages have the correct sign).

Let us assume that we desire $\tilde{A}$ to be the true advantages. Later we discuss why this is a sensible goal. The following result relates different possible error measures .

Theorem Define $|x|^2 =< x, x >$, $\tilde{f}(s) \equiv \sum_a \pi(s,a) f(s,a)$, $J(s) \equiv \sum_a \pi(s,a) Q(s,a)$, and $\tilde{A}(s,a) \equiv f(s,a) - \tilde{f}(s)$. Then

$$|f - Q|^2 = |\tilde{A} - A|^2 + |\tilde{f} - J|^2 \,.$$

proof: This follows from:

$$\begin{aligned} |f - Q|^2 &= |(\tilde{A} - A) - (\tilde{f} - J)|^2 \\ &= |\tilde{A} - A|^2 + |\tilde{f} - J|^2 - 2 < \tilde{A}, \tilde{f} - J > + 2 < A, \tilde{f} - J > \\ &= |\tilde{A} - A|^2 + |\tilde{f} - J|^2 \end{aligned}$$

where we used the fact that A and $\tilde{A}$ have zero expectation values ($\sum_a \tilde{A}(s,a) = 0$ for all s).

Minimizing the l_2 error measure is equivalent to minimizing the angle between the two vectors at hand (assuming that the vector being fit can be scaled). The result shows that the true error in fitting the Q values is an upper bound to the true error on the advantages (which is what is ultimately important to guarantee improvement). If we constrain $\tilde{f} = 0$, then the vector which minimizes the true error with the respect to the state-action values also minimizes the true error with respect to the advantages. However, for a finite sample size this is not necessarily appropriate. Rather than ignoring the baseline, lower error estimates might be obtained if the baseline is fit as well.

Note that for small advantages $|\tilde{A} - A|^2$ will be small but the direction will not necessarily be correct. This is the same issue creeping up again.

4 The Variance Trap

The question we now seek to address is given a direction f how hard is it to determine if this direction will improve the policy? The previous result shows that all that is necessary is to determine the sign of $< f, Q >_\nu$. We will now bound the signal to noise ratio of this estimate.

The related method is to update the policy for small α and check if the policy improves. However, equation 14 shows that this is essentially equivalent to checking if $< f, Q >_\nu$ is positive (i have to think about this a bit more). The large α behavior is not in the incremental regime.

Let us specify a crude model in which we can estimate the signal to noise ratio of $< \tilde{A}, Q >_\nu$. First, let us assume that we can obtain samples from $\nu(s,a) = d(s)\pi(s,a)$. Second, let us assume that the variance of a Monte Carlo estimate, $\chi(s,a)$, of $Q(s,a)$ is $\sigma^2(s,a)$. Now a Monte Carlo estimate of $< \tilde{A}, Q >_\nu$ can be obtained from the sample $\chi(s,a)\tilde{A}(s,a)$ where s, a is sampled under ν. The following theorem gives a bound on the signal to noise ratio of this estimate.

Theorem Assume that samples s, a are sampled from $d(s)\pi(s,a)$ and that the unbiased Monte Carlo estimates, $\chi(s,a)$, of $Q(s,a)$ have variance $\sigma^2(s,a)$. Assume that f is some vector and $\tilde{A}(s,a) \equiv f(s,a) - \sum_{a'} \pi(s,a')f(s,a')$. Then $\chi(s,a)\tilde{A}(s,a)$ is an unbiased estimate of $< \tilde{A}, Q >_\nu$. Further, for any set of states M, let ε be the probability of $s \notin M$, and let $< x >_{s\in X}$ be the conditional expectation $\sum_{s,a} P(s|s \in X)\pi(s,a)x(s,a)$ (where for $s \in X$, $p(s|s \in X) = d(s)/P(s \in X)$). Then the signal to noise ratio of this estimate with respect to any f is bounded by

$$\sigma^2/\mu^2 \geq \frac{1}{\varepsilon < (\frac{A}{Q})^2 >_{s\notin M} + < (\frac{A}{\sigma})^2 >_{s\in M}} - 1\,.$$

Before we provide the proof, let us discuss this result. Consider a set of states M in which the agent is almost always in, say with probability $1 - \varepsilon$ (*eg* states far away from the goal in example 1). The advantages in these states could be very small (in comparison to the noise) making it difficult to improve the policy in these states (*eg* $< (\frac{A}{\sigma})^2 >_{s\in M}$ is exponentially small in example 1*)* . Now consider the states in $\overline{M}$ (the states not in M). These states could have large advantages, but the agent never has access to them due to the fact that the agent does not visit them very often (*eg* ε is exponentially small in example 1). The previous result then states that determining if *any* f will improve the policy could be very difficult, because of very high signal to noise. Thus the agent will have a hard time improving it's policy. Unfortunately, since there could be large advantages in $\overline{M}$ the agent stands to gain a large performance increase (up to $A_\infty/(1-\gamma)$) by improving it's policy.

As shown in section 3.3, gradient methods are just a special case. *For certain polices and MDPs, this lower bound can be arbitrarily large (as in the 2-state MDP) and can be exponentially small with respect to the size of the state space (as in example 1). Since this result holds for all f, then there does not exist any class of parameterized policies for which the gradient can be estimated with low a signal to noise ratio for these aforementioned cases, since the inner product being positive determines being correct to within 90 degrees.* Contrast this to bound 3.

Intuitively, this statement reflects the fact the average reward is a performance measure that is insensitive to changes in the policy at states that are visited infrequently under the stationary distribution.

Proof: The mean μ is:

$$\begin{aligned}
\mu &= \sum_{s,a} d(s)\pi(s,a)\int dP(\chi(s,a))\chi(s,a)\tilde{A}(s,a) \\
&= \sum_{s,a} d(s)\pi(s,a)Q(s,a)\tilde{A}(s,a) \\
&= <\tilde{A}, Q>_\nu
\end{aligned}$$

and the variance is:

$$\begin{aligned}
\sigma^2 &= <\chi^2> - \mu^2 && (11)\\
&= \sum_{s,a} d(s)\pi(s,a)\int dP(\chi(s,a))\chi^2(s,a)\tilde{A}^2(s,a) - \mu^2 && (12)\\
&= \sum_{s,a} d(s)\pi(s,a)\tilde{A}^2(Q^2(s,a)) + \sigma^2(s,a)) - \mu^2\,. && (13)
\end{aligned}$$

Technically, $\tilde{A}$ is constrained such that $\sum_a \tilde{A}(s,a)\pi(s,a) = 0$ (by definition, $\tilde{A}_{sa} \equiv \tilde{f}_{sa} - \sum_{a'} \pi_{sa'}\tilde{f}_{sa'}$). However, we use the unconstrained minima of σ^2/μ^2 to provide a sufficient lower bound. It is straightforward to show that

$$\frac{\partial(\sigma^2/\mu^2)}{\partial \tilde{A}_{sa}} = 2\mu^2\nu(s,a)\tilde{A}(s,a)(Q^2(s,a)+\sigma^2(s,a))+2\nu(s,a)A(s,a)\sum_{s,a}\nu(s,a)\tilde{A}^2(Q^2(s,a))+\sigma^2(s,a))\,.$$

The extrema to this equation is at $\tilde{A}^*(s,a) = \frac{A(s,a)}{Q^2(s,a))+\sigma^2(s,a)}$, and it straightforward to check that this is the minima.

aside: *Intuitively, this last equation for $\tilde{A}^*$ makes sense if you think about it. Consider the 0 noise case. Then, $\tilde{A}^* = 1/A$ so that every sample is equal to one, which has 0 variance.*

By substitution,

$$\sigma^2/\mu^2 \geq (\sum_{s,a}\nu(s,a)A^*(s,a)A(s,a))^{-1} - 1\,.$$

Clearly, $\tilde{A}^*(s,a) \leq \frac{A(s,a)}{\sigma^2(s,a)}$ and $\tilde{A}^*(s,a) \leq \frac{A(s,a)}{Q^2(s,a))}$. Thus,

$$\begin{aligned}
\sum_{s,a}\nu(s,a)A^*(s,a)A(s,a) &= \sum_{s\notin M,a}\nu(s,a)A^*(s,a)A(s,a) + \sum_{s\in M,a}\nu(s,a)A^*(s,a)A(s,a) \\
&\leq \sum_{s\notin M,a}\nu(s,a)\frac{A^2(s,a)}{Q^2(s,a)} + \sum_{s\in M,a}\nu(s,a)\frac{A^2(s,a)}{\sigma^2(s,a)} \\
&= \varepsilon\sum_{s,a}P(s|s\notin M)\pi(s,a)\frac{A^2(s,a)}{Q^2(s,a)} + (1-\varepsilon)\sum_{s\in M,a}P(s|s\in M)\pi(s,a)\frac{A^2(s,a)}{\sigma^2(s,a)} \\
&\geq \varepsilon < (\frac{A}{Q})^2 >_{s\notin M} + < (\frac{A}{\sigma})^2 >_{s\in M}
\end{aligned}$$

where we have used $P(s|s \notin M) = P(s\&s \notin M)/P(s \notin M) = d(s)I_M(s)/\varepsilon$, where indicator function $I_{\overline{M}}(s)$ is 1 if $s \notin M$ else 0 (and the related equation for $P(s|s \in M)$). The result follows from joining these inequalities. **q.e.d.**

5 A Different Performance Measure

This variance trap leads to two problems that we seek to address. The first is that estimating an f that will improve the average reward could be very difficult if the large advantages are at states not visited often. We view this as problem of exploration. The agent must have access to a state to improve the policy at that state. Second, given any f it can be difficult to do a line search to check if the policy improves due to the fact that the policy could be improving at unlikely states, but not at likely states. For instance, in example 1, assume that at the states closest to the goal the agent has fit the advantages (perhaps by using some exploration scheme) but has not accurately approximated the advantages at the states far away from the goal states (due to the low signal to noise of the advantages at these states). A line search checking if the average reward improves would not be feasible since improvement at the policy near the goal states will not be detected as increasing the current average reward (though it does slightly increase).

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. ***The assumption we make in this paper is that the agent has access to a simulator that can place the agent in any state.***

First, let us consider the recurrent state formulation. In this formulation, we assume that for every policy there exists one unique recurrent state which the agent will always return to (*eg* in example 1, we could choose the left most state to be the recurrent state). The average reward can then be estimated by the reward obtained in between visits to this recurrent state divided by the average time to return to the recurrent state. In this formulation, the state-action values and state values are defined as the reward obtained until this recurrent state is reached. We can define a different performance ω as

$$\omega \equiv \sum_s m(s) J^\pi(s)$$

where $m(s)$ is some fixed distribution (possibly uniform). This has the effect that agent will be maximizing the reward from each state to the recurrent state. Thus even if a state is not visited often (like the right most states in example 1), this performance measure will give significant weight to improve the reward from this unlikely state. In general, an improvement of ω will not necessarily improve the average reward. However, clearly the maximum of ω is a maximum of η, since the maximum of ω will maximize the undiscounted reward from each state simultaneously which also maximizes the average reward. We will return to this point in the next section and discuss how to improve both performance measures.

5.1 Incremental Improvement

In general, the discounted formulation might be problematic since there does not necessarily exist a common recurrent state for every policy (or the time to the recurrent state could be very long even if the process mixes quickly). Instead, we consider using the discounted reward formulation. Using a discount factor is an approximation resting on the fact that for large γ each discounted reward $(1-\gamma)J_\gamma(s) \approx \eta$. As argued in ??, to approximately maximize the average reward it is sufficient to set $\frac{1}{1-\gamma}$ to be the order of the mixing time.

aside: for now i am still leaving in un-normalized quantities. the reason is that it is the un-normalized quantities that appear in the gradients for the discounted and undiscounted formulations (see equations 16 and 6). I'll probably change it later, but for now it's also important to keep in mind this fact. Also there is no such concept of a normalized advantage in the undiscounted case!

First, note that the average reward is equal to expected discounted reward [6] (where the expectation is with respect to the stationary distribution), or:

$$\eta(\pi) = (1-\gamma)\sum d^\pi(s) J^\pi_\gamma(s)\,.$$

This follows from using the consistency condition ($J = r + \gamma P J$):

$$\begin{aligned}\sum_s d^\pi(s)J(s) &= \sum_s d^\pi(s) r + \gamma \sum_{s,s'} d^\pi(s) P_{s\to s'} J(s') \\ &= \eta + \gamma \sum_{s'} d^\pi(s') J(s')\end{aligned}$$

where we have also used the stationarity of $d(s)$. Collecting terms proves the previous statement. This motivates the following new performance measure:

$$\omega = (1-\gamma)\sum_s m(s) J^\pi_\gamma(s)$$

Maximizing this performance measure is equivalent to maximizing the discounted reward from starting states chosen under the distribution $m(s)$ (which could be uniform). This alternative performance measure is analogous to the undiscounted one given above.

Now let us analyze the improvement conditions for this new performance measure and its signal to noise properties. Given a start state s_0, let us define the discounted future distribution of states as:

$$d^\pi_\gamma(s|s_0) \equiv (1-\gamma)\sum_{\tau=0}^{\infty}\sum_s \gamma^\tau P(s_\tau = s|s_0)$$

, which is normalized to one. This definition is analogous to the stationary distribution in the undiscounted setting. Since the starting state is chosen according to $m(s)$, the following definition will be useful:

$$\tilde{m}(s) \equiv \sum_{s'} d_\gamma^\pi(s|s')m(s') .$$

Also, we will use the following inner product:

$$< x, x' >_{\tilde{m}\pi} = \sum_{s,a} \tilde{m}(s)\pi(s,a)x(s,a)x'(s,a) .$$

Theorem Assume the definition given above for $\tilde{m}(s)$. Let us consider the update rule $\pi_s(a) \leftarrow \pi_s(a)(1+\alpha f(s,a))$ for some vector f and some small α. Let A be the true advantages. Then the performance improvement of ω is:

$$\Delta\omega = \alpha < A, f >_{\tilde{m}\pi} + O(\alpha^2) . \tag{14}$$

Proof The proof is similar to the previous related theorem (don't know how to make a real Theorem env. in lyx so can't label them. arrgh.)

The gradient of $J_\gamma(s_0)$ is

$$\begin{aligned}(1-\gamma)\nabla J_\gamma(s_0) &= (1-\gamma) \sum_{\tau=0}^{\infty}\sum_s \gamma^\tau P(s_\tau = s|s_o)\sum_a \nabla\pi(s,a)Q_\gamma(s,a) &(15)\\ &= \sum_s d_\gamma(s|s_0)\sum_a \nabla\pi(s,a)Q_\gamma(s,a) &(16)\end{aligned}$$

Let us consider the table-lookup Gibbs distribution, $\tilde{\Pi}_G = \{\pi_\theta : \pi(s,a) \propto \exp(\theta_{sa})\}$. This gradient is then:

$$(1-\gamma)\partial_{sa}J_\gamma(s_0) = d_\gamma(s|s_0)\pi(s,a)A_\gamma(s,a)$$

where $A_\gamma(s,a) \equiv Q_\gamma(s,a) - J_\gamma(s)$ and $\partial_{sa} \equiv \partial/\partial\theta_{sa}$. It follows that the gradient of ω is :

$$\begin{aligned}\partial_{sa}\omega(\pi) &= (1-\gamma)\sum_{s'} m(s')\partial_{sa}J_\gamma(s') &(17)\\ &= \sum_{s'} m(s')d_\gamma(s|s')\pi(s,a)A_\gamma(s,a) &(18)\\ &= \tilde{m}(s)\pi(s,a)A_\gamma(s,a). &(19)\end{aligned}$$

Similar arguments as before lead to $\Delta\omega = \sum_{sa}\partial_{sa}\omega\Delta\theta_{sa}$ (for small $\Delta\theta$) and so:

$$\Delta\omega = \sum_{sa}\tilde{m}(s)\pi(s,a)A(s,a)(\alpha\tilde{f}_{sa}) = < A, \tilde{f} >_{\tilde{m}\pi} .$$

q.e.d.

Also, if this is the performance measure that is being used then line searches should check if ω increases.

5.2 Lower Bounding The Signal to Noise

Clearly, both ω and η will be maximized by the globally optimal policy. However, previously we argued that if we only are concerned with improving η then we could be trapped in at a policy where determining if a direction will increase the policy could have an arbitrarily low signal to noise ratio, eventhough there are significant performance improvements to be had. We now show that under ω, the signal to noise in determining if a direction will improve the policy can only become arbitrarily small when the policy has performance close to the globally optimal policy.

First, let us obtain a better handle on what the performance measure is optimizing. Although for large γ each discounted reward $(1-\gamma)J_\gamma(s) \approx \eta$, the difference will be some be some transient $\tau(s)$, such that $J_\gamma(s) = \eta/(1-\gamma) + \tau(s)$. This equation $J_\gamma(s) = \eta/(1-\gamma) + \tau(s)$ shows that there are two terms contributing to the discounted reward. The first term shows that the increase in the average reward will be seen $1/(1-\gamma)$ times. However, even with this scaling, $\Delta\eta/(1-\gamma)$ could be arbitrarily small under policy improvement at a state with a low stationary distribution. To put it another way, the mixing time ($\approx 1/(1-\gamma)$) could much greater than the time it takes to get a sample of state under the stationary distribution (as in example 1 and 2). In contrast, the second term *always* is impacted by the transient reward until the states have mixed. In terms of these transients:

$$\omega = (1-\gamma)(\sum_s \tau(s) + \eta/(1-\gamma)) .$$

In example 1, the rewards incurred till the process has mixed will be observed in the transients, including the transients starting from the unlikely states (whose improvement could have little impact on $\eta/(1-\gamma)$).

***aside*:** *The connection to the recurrent case is that the first term in $\eta/(1-\gamma)+\tau(s)$ is essentially the infinite return after the recurrent state is reached (and can be ignored since it is the same for all states). Thus $\tau(s)$ is analogous to the reward until a recurrent state is reached. It is a form of additive normalization where $\eta/(1-\gamma)$ is subtracted from every state (as $\gamma \to 1$). This might be obvious to you but it still confuses me at times.*

The following theorem shows that there exists a direction which improvement of ω can always be detected, unless the policy has performance close to the globally optimal policy (with the caveat that we never let the policy become arbitrarily deterministic).

Theorem Assume that Assume that samples s, a are sampled from $\tilde{m}(s)\pi(s,a)$ and that the unbiased Monte Carlo estimates, $\chi(s,a)$, of $Q(s,a)$ have variance $\sigma^2(s,a)$. Assume that $\max_{s,a}\sigma(s,a) = \sigma_\infty$, $\max_{s,a}|A(s,a)| = A_\infty$, $\min_{s,a}\pi(s,a) = \pi_0$, and $\min_s m(s) = m_0$. Let the improvement direction f be the true advantages A. Then the estimate $\chi(s,a)\tilde{A}(s,a)$ of $< \tilde{A}, Q >_{\tilde{m}\pi}$ will have the following bounded noise to signal ratio:

$$\sigma^2/\mu^2 \leq \frac{\sigma_\infty^2}{(1-\gamma)\pi_0 m_0 A_\infty^2} .$$

Further, the if η^* is the optimal possible average reward, then $|\eta - \eta^*| \leq A_\infty/(1-\gamma)$.

Note that for the uniform distribution for $m(s)$ that $m_0 = 1/(\#\text{ of states})$, whereas in the previous bound ε could be exponentially small (in terms of the size of the state space). Also, although this bound uses π_0 (the lowest action probability of the *current* policy), this could be avoided by using a starting state-action distribution, $m(s,a)$, rather than just a starting-state distribution $m(s)$. *I'll change this later, not sure why I didn't use m(s,a) to start with.*

proof: Using $f = A$, the mean is

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

and using equation 13, the variance of an estimate can be bounded by

$$\begin{aligned}
\sigma^2 &\leq \sum_{s,a} \tilde{m}(s)\pi(s,a)\sigma^2(s,a)A^2(s,a) \\
&\leq \sigma_\infty^2 \sum_{s,a} \tilde{m}(s)\pi(s,a)A^2(s,a) \\
&= \sigma_\infty^2 \mu \,.
\end{aligned}$$

Thus,

$$\sigma^2/\mu^2 \leq \sigma_\infty^2/\mu \,.$$

Using the definition of d_γ, $d(s|s_0) \geq (1-\gamma)$ so $\tilde{m}(s) \geq d_\gamma(s|s)m(s) \geq (1-\gamma)m_0$. Also, let s_∞ be the state with the largest advantage. It follows that:

$$\begin{aligned}
\mu &= \sum_{s,a} \tilde{m}(s)\pi(s,a)A^2(s,a) \\
&\geq \tilde{m}(s_\infty)\pi(s_\infty,a)A^2(s_\infty,a) \\
&\geq (1-\gamma)\pi_0 m_0 A_\infty^2 \,.
\end{aligned}$$

It is straightforward to show if the maximum advantage is A_∞ then the global policy can be no greater then $A_\infty/(1-\gamma)$ than the current policy (and this proof is omitted for now). **q.e.d.**

6 Have we thrown the baby out with the bath water?

This section is bit rough.

An important question raised by this analysis is if we try improve ω could the actual average reward decrease (*which is what we wanted to maximize in the first place!*). Note that policy iteration simultaneously improves both these measures on any policy improvement step (because the future reward from every state is increased). In general, an improvement of ω could decrease η. However, if f induces an $\tilde{A}$ such that for all states and actions the sign of $\tilde{A}(s,a)$ is the same as that of $A(s,a)$ then both performance measures will improve. The weaker condition is that both $< Q, f >_{\tilde{m}\pi}$ and $< Q, f >_{d\pi}$ (where $d(s)$ is the stationary distribution) must be positive for both performance measures to improve.

When choosing f to improve ω, we can place some additional pressure to try to force $< Q, f >_{d\pi}$ to be positive as well. Alternatively, we could just force the l_2 error, $< \tilde{A} - A, \tilde{A} - A >_{\tilde{m}\pi}$ (or $< f - Q, f - Q >_{\tilde{m}\pi}$), to be sufficiently small such that $< \tilde{A} - A, \tilde{A} - A >_{d\pi}$ (or $< f - Q, f - Q >_{d\pi}$) is small as well. We definitely can improve ω (as long are there are some large advantages somewhere, else we are close to global optimality). However, even if $< \tilde{A} - A, \tilde{A} - A >_{d\pi}$ is forced to be small, this does not guarantee improvement of η, since the error could be small if all the advantages at states that are visited often are small. This is essentially the signal to noise argument revisited (*ie* we know that in general it can be very difficult to force η to increase). However, looking on the bright side, if we cannot detect if η increases, we will not be able to detect a decrease in the average reward either (for exactly the same reasons).

The punchline is that we can fit the $\tilde{A}$ good enough under ω, such that either: 1) ω increases or 2) if we cannot detect improvement of ω, then the policy must have near optimal performance. *Further, while this is being done, we can also ensure that we never detect (with some certainty) a decrease in the average reward.* Thus, when doing line searches we have to check that *both* ω and η are increasing or at least non-decreasing (where the noise determines the quality of our checks). For instance, in example 1, the policy could improve close to the goal state which would increase ω (and we could detect this increase!). However, the improvement to the average reward would be undetectable when the policy is changed at these highly unlikely states. Eventually (when the states close to goal have a good policy), the states visited with reasonable probability under the stationary distribution will have large advantages, and then improvement will be seen in both performance measures.

So it seems like we can have our cake and eat it too.

References

[1] P. Bartlett and J. Baxter. Estimation and approximation bounds for gradient-based reinforcement learning. Technical report, Australian National University, 2000.

[2] M. Kearns and S. Singh. Near-optimal reinforcement learning in polynomial time. *Proc. 15th International Conf. on Machine Learning*, 1998.

[3] P. Marbach and J. Tsitsiklis. Simulation-based optimization of markov reward processes. Technical report, Massachusetts Institute of Technology, 1998.

[4] L. Peshkin an K. Kim N. Meuleau. Exploration in gradient-based reinforcement learning. Technical report, Massachusetts Institute of Technology, 2001.

[5] D. Precup, R. Sutton, and S. Dasgupta. Off-policy temporal-difference learning with function approximation, 2001.

[6] S. Singh, T. Jaakkola, and M. I. Jordan. Learning without state-estimation in partially observable markovian decision processes. *Proc. 11th International Conference on Machine Learning*, 1994.

[7] R. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. *Neural Information Processing Systems*, 13, 2000.

[8] S. B. Thrun. Efficient exploration in reinforcement learning. Technical report, Carnegie Mellon University, 1992.

[9] S. D. Whitehead. Complexity and cooperation in q-learning, 1991.