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. Maximizing this alternative performance measure 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 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 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.

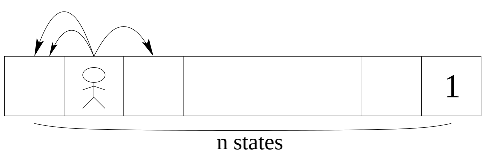

Figure 1.1. Example 1: Two actions move agent to the left and one actions moves agent to the right.

The average reward is what we ultimately seek to maximize. However, this performance measure has the fundamental problem that it is insensitive to performance improvement at highly unlikely states. Improvement at these states can lead to drastic policy improvement (as in example 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.

(1) In section 2, the problems with gradient methods are exhibited in more detail.
(2) In section 3, we discuss general incremental policy improvement (and consider both gradient methods and function approximation).
(3) In section 4, we show that determining if any direction (in policy space) will improve the average reward can be arbitrarily difficult.
(4) In section 5, we propose an alternative performance measure 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.
(5) In section 6, we discuss how we can improve this alternative performance measure without ever detecting a decrease in the average reward.

2. Preliminaries

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) with rewards. An MDP with rewards 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. 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 distribution d_π. Under this assumption, the *average reward* is $\eta_\pi \equiv \sum_{s,a} d_\pi(s)\pi(a|s)R(s,a)$, the state-action value is

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

and the value function is

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

, where and s_t and a_t are the state and action at time t. One last definition is required: The advantage is defined by:

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

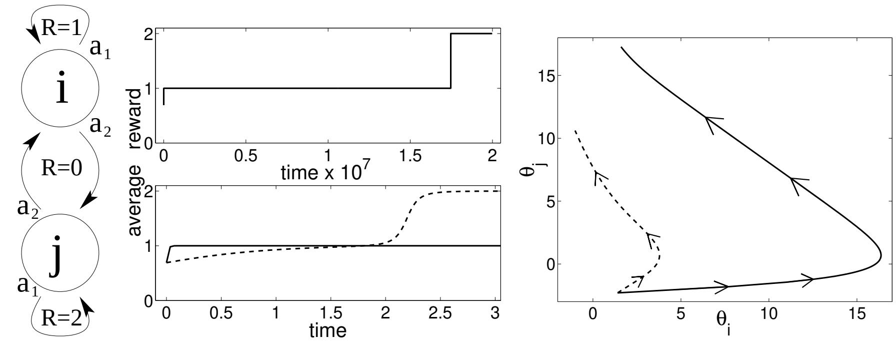

FIGURE 3.1. Example 2: A) See text. B top) The average reward vs. time (on a 10^7 scale) of a policy under standard gradient descent using the sigmoidal policy parameterization ($\pi(1; s, \theta_i) \propto \exp(\theta_i)/(1+\exp(\theta_i))$), with the initial conditions $\pi(i, 1) = .8$ and $\pi(j,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

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

3. What is Wrong with Policy Gradient?

This section first shows another example of how policy gradient techniques can behave poorly, then discusses criteria for policy gradient improvent and finishes with a theorem stating that policy gradient techniques can converge remarkably slowly.

3.1. **Examples.** In this section, we show an example exhibiting how the policy gradient approach can go drastically wrong. Later, we will formalize how policy gradient approaches can go wrong.

The introductor example in figure 1.1 shows that Policy Gradient techniques can be relatively slow to converge. However, the situation is (perhaps) even worse than might thought. In fact, policy gradient techniques can behave very poorly for even a 2 state problem. Our example will use the Gibbs table-lookup distribution, $\tilde{\Pi}_G = \{\pi_\theta : \pi(a|s) \propto \exp(\theta_{sa})\}$. The exact gradient of the average reward is then:

$$\nabla\eta(\pi) = \sum_{s,a} d_\pi(s)\nabla\pi(a|s)Q_\pi(s,a) \tag{3.1}$$

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

$$\partial_{sa}\eta(\pi) = d_\pi(s)\pi(a|s)(Q_\pi(s,a) - J_\pi(s)) = d_\pi(s)\pi(a|s)A_\pi(s,a) \tag{3.2}$$

Consider the 2-state MDP shown in Figure 3.1B. In this MDP, 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 squelches the learning at state j and this leads 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 6.

This example is remakably 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.

3.2. **Conditions for Policy Improvement.** Gradient methods guarantee improvement due to the incremental nature of the algorithm although no *rate* of improvement guarantees are made. Performance guarantees exist only for algorithms which change the policy incrementally (by following a gradient or via value function approximations).

Let's now consider the incremental update rule:

$$\pi(a|s) \propto \pi(a|s)(1 + \alpha f(s,a)) \tag{3.3}$$

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.

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

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

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

Theorem. *Let $A_\pi(s,a)$ be the advantages of any state and action. For all f, there exists an ϵ s.t. for all $\alpha < \epsilon$ the update rule (equation 3.3) has performance improvement:*

$$\Delta\eta = \alpha < A_\pi, f >_{d\pi} + O(\alpha^2) \tag{3.4}$$

Proof. Specialization of 4.1 which is proved later. □

This theorem state 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 $\tilde{A}(s,a) \equiv f(s,a) - \sum_{a'} \pi(a'|s) f(s,a')$ then we get:

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

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 $< Q, \tilde{A} >_{d\pi}$ since the state-action values can be estimated from Monte-Carlo samples using the empirically observed average reward.

In light of the last theorem, how do gradient method techniques perform? 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 $\langle \partial \log \pi/\partial\theta, Q\rangle_\nu > 0$ 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 6.1) can be written in the form:

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

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 the multiple parameter case, the parameterization will choose a set of directions, $f_i = \partial \log \pi/\partial\theta_i$, and the gradient will tell us to move in the f_i or $-f_i$ direction with respect to each parameter θ_i. 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.

3.3. **The Variance Trap.** Given a direction f how hard is it to determine if this direction will improve the policy? We will now bound the signal to noise ratio of this estimate of $< f, Q >_\nu$.

A related method is to perform a line search to determine if the policy improves. Equation 4.1 shows that this is essentially equivalent to checking if $< f, Q >_\nu$ is positive for small changes to the policy.

Under a simple sampling model, we can bound on the signal to noise ratio of an estimate of $< \tilde{A}, Q >_{d\pi}$.

Theorem 3.1. *Assume that samples s, a are sampled from $d(s)\pi(a|s)$ 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 $\tilde{A}(s,a) \equiv f(s,a) - \sum_{a'} \pi(a'|s) f(s,a')$. Then $\chi(s,a)\tilde{A}(s,a)$ is an unbiased estimate of $< \tilde{A}, Q >_\nu$ and the noise to signal ratio with respect to any f is bounded by:*

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

Proof. Since the χ are unbiased, the mean μ is:

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

and the variance is:

$$\Sigma^2 = \left\langle \chi^2 \right\rangle - \mu^2 \tag{3.5}$$

$$= \sum_{s,a} d(s)\pi(a|s) \int dP(\chi(s,a))\chi^2(s,a)\tilde{A}^2(s,a) - \mu^2 \tag{3.6}$$

$$= \sum_{s,a} d(s)\pi(a|s)\tilde{A}^2(Q^2(s,a) + \sigma^2(s,a)) - \mu^2 \,. \tag{3.7}$$

giving us:

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

Technically, $\tilde{A}$ is constrained such that $\sum_a \tilde{A}(s,a)\pi(a|s) = 0$. However, we use the unconstrained minima of Σ^2/μ^2 to provide a lower bound.

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

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

The extrema to this equation is at $\tilde{A}^*(s,a) = \frac{Q(s,a)\left\langle \tilde{A}^2(Q^2+\sigma^2) \right\rangle_{d\pi}}{\mu(Q^2(s,a))+\sigma^2(s,a))}$

I'm lost here. How do you get to the next step? -John

begin answer: woops, I left out some stuff here. I wanted something quick and dirty, and the following manipulations worked. Technically, $\tilde{A}$ is constrained such that $\sum_a \tilde{A}(s,a)\pi(a|s) = 0$ (or equivalently $\tilde{A}(s,a) \equiv f(s,a) - \sum_{a'} \pi(a'|s)f(s,a')$). So, it is easy to show:

$$\mu = \left\langle \tilde{A}, A \right\rangle_{d\pi} \,.$$

And so:

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

The goal is minimize $\frac{\Sigma^2}{\mu^2}$ subject to the aforementioned constraint. However, lets just minimize this quantity unconstrained since that gives a bound. In essence, this method sort of uses the constraint for μ but ignores it for Σ. so:

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

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

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

$$\frac{2d(s)\pi(a|s)\tilde{A}^*(s,a)(Q^2(s,a)+\sigma^2(s,a))}{\mu^2} = \frac{2d(s)\pi(a|s)A(s,a)}{\mu^2}$$

and the second term is

$$\begin{aligned}\frac{2d(s)\pi(a|s)A(s,a)\left\langle \tilde{A}^{*2}(Q^2+\sigma^2)\right\rangle_{d\pi}}{\mu^3} &= \frac{2d(s)\pi(a|s)A(s,a)\left\langle \tilde{A}^* A\right\rangle_{d\pi}}{\mu^3} \\ &= \frac{2d(s)\pi(a|s)A(s,a)}{\mu^2}\end{aligned}$$

where we have substituted the expression for $\tilde{A}$ in just one of the terms in $\tilde{A}^2$.

This might not be the cleanest method, but I knew roughly what I wanted, and I was flailing doing it exactly with Lagrange multipliers. This gives a good enough somewhat clean result, which is good enough for now.

-Sham

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

$$\Sigma(\tilde{A}^*)^2 = \mu(\tilde{A}^*) - \mu(\tilde{A}^*)^2\,. \tag{3.8}$$

The result follows by using $\Sigma^2/\mu^2 \geq 1/\mu(\tilde{A}^*) - 1$. □

Let us discuss why this ratio could be large. 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). If the advantages of the states in M are small (as compared to the noise level) then these states will have a negligible contribution to the signal to noise. Further, if there are large advantages at states not in M, then the contribution these states have to signal to noise will be down-weighted by ε (ε is exponentially small in example 1). Thus the agent will have a hard time improving it's policy. Unfortunately, since there could be large advantages in the states not in M, the agent stands to gain a large performance increase.

Corollary 3.2. *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 $\sum_{s,a} P(s|s\in X)\pi(a|s)x(s,a)$ (where for $s \in X$, $p(s|s\in X) = d(s)/P(s\in X)$). Then:*

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

Proof. Assume $\tilde{A}^*(s,a) = \frac{A(s,a)}{Q^2(s,a))+\sigma^2(s,a)}$, as before. Using $\tilde{A}^*(s,a) \le \frac{A(s,a)}{\sigma^2(s,a)}$ and $\tilde{A}^*(s,a) \le \frac{A(s,a)}{Q^2(s,a))}$,

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

The result follows from 3.8. □

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 6.3. It should be straight forward to show that this signal to noise ratio scales exponentially in example 1.

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.

4. 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.

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

A natural candidate for a performance measure is to weight the improvement from all states more uniformly, such as

$$\omega \equiv \sum_s m(s) * (\text{future reward from state } s) \, .$$

Under this performance measure, even if a state is not visited often (like the right most states in example 1), importance will be given to improving the future reward from this unlikely state. Note that the policy that maximizes ω also simultaneously maximizes the future reward from every state.

The problem with directly using this measure as stated is that the sum future reward from any state s is infinite. The dynamic programming formulation for the undiscounted, infinite horizon scenario is to solve the following, slightly modified, Bellman equations:

$$J^*(s) = \max_a (r(s,a) - \eta^* + \sum_{s'} p(s'|s,a) J^*(s'))$$

for both J^* and η^*. The optimal policy is then $\pi^*(a|s) = \text{argmax}_a (r(s,a) - \eta(\pi) + \sum_{s'} p(s'|s,a) J^*(s'))$. Note that if $J^*(s)$ is a solution then so is $J^*(s)$ plus a constant, and so there is a whole family of solutions. The differential value functions defined by $J_\pi(s) \equiv E_\pi\{\sum_{t=0}^{\infty} (R(s_t) - \eta_\pi) | s_0 = s\}$, for the policy π^*, are one such solution. A common alternative is to define the value functions by the reward received until a certain recurrent state (so $J_\pi(s) \leftarrow J_\pi(s) - J_\pi(s_{\text{recurrent}})$). All these suggestions satisfy the consistency condition $J_\pi(s) = r(s,a) - c + \sum_{s'} p(s'|s,a) J_\pi(s')$ for some constant c.

The problem is that we cannot use the performance measure $\omega \equiv \sum_s m(s) J_\pi(s)$ for any of these suggestions. *The reason is that J^* is not maximized when the undiscounted, infinite horizon Bellman equations are solved, unlike in the finite horizon or discounted case.* Intuitively, the $J^*(s)$ is the transient after an infinite amount of reward has been subtracted away ($J_\pi(s) = E_\pi\{\sum_{t=0}^{\infty} (R(s_t) - \eta_\pi) | s_0 = s\} \approx E_\pi\{\sum_{t=0}^{\infty} R(s_t) | s_0 = s\} - \infty \eta^*$. It is η^* that is being maximized not these differential value functions. **aside: This is an important point that we *cannot* define an 'exact' alternative performance measure ω in the undiscounted, infinite horizon case. My previous draft was incorrect in suggesting that the recurrent state formulation would work! I have not thought through all the implications of this yet.**

To overcome these difficulties, we no longer considering maximizing the undiscounted, future reward, but an approximation to it. There are two sensible alternatives, using the T-step future reward or using γ discounted reward, either of which is acceptable. For computational and notational convenience, we only consider the discounted formulation.

aside: In the T-step scenario, we have to not only consider the T-step future reward, but also have to the $T-1, T-2, \ldots 1$-step future rewards. For instance, just look at $\nabla J_T(s_0) = \sum_{t=1}^{T} \sum_s P(s_t = s|s_0) \nabla \pi(a|s) Q_{T-t}(s)$, where Q_{T-t} is the reward for the next $T-t$ steps. Nothing is wrong with this, but the inner-product condition is not as clean due to an additional sum over 1 to T.

We now consider the following performance measure:

$$\omega(\pi) = (1-\gamma) \sum_s m(s) J_\gamma(s)$$

Maximizing this performance measure is equivalent to maximizing the discounted reward from starting states chosen under the distribution $m(s)$ (which could be uniform). Unlike in the undiscounted, infinite horizon setting, the $J_\pi(s)$ can be *simultaneously maximized* when the corresponding Bellman equations, $J_\gamma(s) = \max_a (r(s,a) + \gamma \sum_{s'} p(s'|s,a) J_\gamma(s'))$, are solved.

Note that if $m(s)$ is equal to the stationary distribution of π then $\omega(\pi)$ will be equal to the average reward of π (of course $m(s)$ is constant where as $d_\pi(s)$ is a function of the policy). More formally (as shown in [?]):

$$\eta(\pi) = (1-\gamma) \sum d_\pi(s) J_\pi(s) \,.$$

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

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

where we have also used the stationarity of $d(s)$. Collecting terms proves the previous statement.

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:

$$d_\pi(s|s_0) \equiv (1-\gamma)\sum_{\tau=0}^{\infty}\sum_s \gamma^\tau P(s_\tau = s|s_0)\,.$$

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:

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

Theorem 4.1. *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:*

$$\Delta\omega = \alpha < A, f >_{\tilde{m}\pi} + O(\alpha^2)\,. \tag{4.1}$$

Proof. The proof is similar to that of theorem 2.1. The gradient of $J_\gamma(s_0)$ is

$$(1-\gamma)\nabla J_\gamma(s_0) = (1-\gamma)\sum_{\tau=0}^{\infty}\sum_s \gamma^\tau P(s_\tau = s|s_o)\sum_a \nabla\pi(a|s)Q_\gamma(s,a) \tag{4.2}$$

$$= \sum_s d_\gamma(s|s_0)\sum_a \nabla\pi(a|s)Q_\gamma(s,a) \tag{4.3}$$

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

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

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 :

$$\partial_{sa}\omega(\pi) = (1-\gamma)\sum_{s'} m(s')\partial_{sa}J_\gamma(s') \tag{4.4}$$

$$= \sum_{s'} m(s')d_\gamma(s|s')\pi(a|s)A_\gamma(s,a) \tag{4.5}$$

$$= \tilde{m}(s)\pi(a|s)A_\gamma(s,a). \tag{4.6}$$

Similar arguments as before lead to the result. □

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 (with the caveat that we never let the policy become arbitrarily deterministic).

Theorem 4.2. *Assume that samples s, a are sampled from $\tilde{m}(s)\pi(a|s)$ and that the unbiased Monte Carlo estimates, $\chi(s,a)$, of $Q(s,a)$ have variance $\sigma^2(s,a)$. Assume that* $\max_{s,a}\sigma(s,a) = \sigma_\infty$, $\max_{s,a}|A(s,a)| = A_\infty$, $\min_{s,a}\pi(a|s) = \pi_0$, *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). *aside: 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

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

and using equation 3.7, the variance of an estimate can be bounded by

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

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:

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

It is straightforward to show if the maximum advantage is A_∞ then the optimal payoff can be no greater then $A_\infty/(1-\gamma)$ than the current policy (this proof is omitted for now). □

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

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 (with the caveat that the γ must be large enough). In general, an improvement of ω could decrease η. Of course if both $<Q, f>_{\tilde{m}\pi}$ and $<Q, f>_{d\pi}$ are positive then both performance measures improve.

We could force the l_2 error $< \tilde{A}-A, \tilde{A}-A >_{\tilde{m}\pi}$ to be sufficiently small such that $< \tilde{A}-A, \tilde{A}-A >_{d\pi}$ is small as well (this is analogous to a supervised learning problem where we fit under one distribution and test under another). It should be straight forward to show that if the error $< \tilde{A}-A, \tilde{A}-A >_{\tilde{m}\pi}$ is always smaller than some reasonable, *policy-independent* constant then either we can detect improvement of ω with a line search *or* if cannot detect improvement, then the policy is close to optimal. Unfortunately, we will not be able to say the same for η if $< \tilde{A}-A, \tilde{A}-A >_{d\pi}$ is smaller than some *policy-independent* constant. The problem is that $< \tilde{A}-A, \tilde{A}-A >_{d\pi}$ could be small due to the agent not visiting states with large advantages. This is essentially the signal to noise argument revisited.

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.

aside: You asked me for an example in which ω could increase but η could decrease. Here is an *outline* of one using an MDP similar to that of example 1 (where the agent could collect rewards along the way to state 50). Consider that at state 49 we learn some action to get considerably more reward. However, say at state 35 we don't fit things good enough and we get a slight decrease in the reward from this state. Under ω, the increase at state 49 is offset by the decrease at state 35. However, under d_π, the agent visits 35 much more often than 49, so the agents average reward could decrease. *On the bright side, although the true average reward might drop, we certainly won't be able to detect this drop, since both states are visited so infrequently.* Obviously, if we fit everything perfectly these are not concerns. Realistically, what we desire is that if there are some large advantages at states visited by the agent with reasonable stationary probability, then we should fit these advantages good enough such that η also improves in addition to ω. Does that make sense?

References

6. 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:

$$\nabla\eta(\pi) = \sum_{s,a} d_\pi(s)\nabla\pi(a|s)Q_\pi(s,a) \tag{6.1}$$

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

$$\partial_{sa}\eta(\pi) = d_\pi(s)\pi(a|s)(Q_\pi(s,a) - J_\pi(s)) = d_\pi(s)\pi(a|s)A_\pi(s,a) \tag{6.2}$$

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 $\theta_t = \theta_0 + \int_0^t \nabla\eta^{\pi_t}dt$. *aside: 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 $d(s)$.*

Theorem 6.1. *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 $\tilde{a} \in \mathrm{argmax}_a Q^{\pi_\infty}(s,a)$, $\partial_{sa}\eta^t$ 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 $\tilde{a} \notin \mathrm{argmax}_a Q^{\pi_\infty}(s,a)$ 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 6.2 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 $A^\infty(s) \neq 0$. $\square$

This result 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 large, as is demonstrated in the two state example (*aside: i will make a figure showing how the variance of the grad. estimate drops).* 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).

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

$$Pr(|\nabla\eta - \nabla\eta_{\mathrm{MC}}|_\infty \geq \varepsilon) \geq \frac{Nd}{2}\exp(-s/\tau) + 2sd\exp(-\varepsilon^2 N\frac{c}{s}) \tag{6.3}$$

where $\nabla\eta$ and $\nabla\eta_{MC}$ are the true and sampled estimates of the gradient, s is any number less than N, and c is a constant. The norm $|\nabla\eta - \nabla\eta_{MC}|_\infty$ is the largest difference in the gradient over all the parameters, *ie* $\mathrm{argmax}_{\theta_i}|\frac{\partial\eta}{\partial\theta_i} - (\frac{\partial\eta}{\partial\theta_i})_{MC}|$. The important point is that the estimated direction needs to point within 90 degrees of the true gradient for local improvement to be guaranteed. The bound shows that we can easily force correctness to within ε. However, if the magnitude of the gradient is $O(\varepsilon)$, then an ε infinity error does not imply a correct direction. This example shows that the magnitude of the gradient can be small far away from the optima.

7. Function Approximation

An alternative to policy gradient techniques is using 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?" If each $\tilde{A}(s,a)$ has the same sign as $A(s,a)$ then improvement will be guaranteed. Motivated by this, let us assume that we desire $\tilde{A}$ to be the true advantages. The following theorem decomposes the approximation error of some function f into two components: an action independent component and a component due to advantages.

Theorem 7.1. *Define* $\tilde{f}(s) \equiv \sum_a \pi(a|s)f(s,a)$, *and* $\tilde{A}(s,a) \equiv f(s,a) - \tilde{f}(s)$. *Then*

$$|f - Q_\pi|^2 = |\tilde{A} - A_\pi|^2 + |J - \bar{f}|\,.$$

Proof. This follows from:

$$\begin{aligned} |f-Q|^2 &= |f - \bar{f} - (Q-J) - (J-\bar{f})|^2 \\ &= |(\tilde{A} - A) - (J - \tilde{f})|^2 \\ &= |\tilde{A} - A|^2 + |J - \bar{f}|^2 - 2\left\langle \tilde{A}, J - \bar{f}\right\rangle + 2\left\langle A, J - \tilde{f}\right\rangle \\ &= |\tilde{A} - A|^2 + |J - \bar{f}|^2 \end{aligned}$$

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

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). 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 constraint is not necessarily appropriate. *The issue of how to fit f using baselines merits further discussion.*

I think the point I was considering is how can a fixed architecture FA ever decrease the policy (under a local search), whereas gradient methods, with fixed architecture, always increase the policy. The reason is a baseline issue (gradients essentially fit advantages by constraining $\tilde{f} = 0$). Discussing this should be easy and belongs in the non-parametric section. Although, we are not interested in fixed architecture FA, a sentence or two on the this case would help with insight and tie with previous work done by others. -Sham

Ok. Preserving for later. -John