IEOR 8100: Reinforcement learning

Lecture 6: Provable guarantees for policy gradient methods

By Shipra Agrawal

Based on [Kakade and Langford, 2002] and Schulman et al. [2015].

1 Examples demonstrating problems with the policy gradient algorithm

Below are two examples of situations when estimating the policy gradient direction is difficult. In particular, the lack of exploration in the policy gradient method translates into large number of samples in order to accurately estimate the gradient direction.

1.1 Example 1

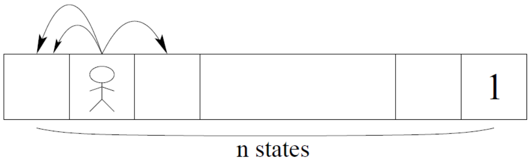

Figure 1: MDP for Example 1

This example from Kakade and Langford [2002] illustrates a scenario where non-zero estimates of policy gradient require observing sample trajectories of exponential length.

We are given an MDP as in Figure 1.1. Each state $s \in \{n, n-1, \dots, 1\}$ has 3 possible actions. For two of those actions, the next state is deterministically $s+1$, and for one action it is deterministically $s-1$. The reward for all actions is 0 in all states except state 1 where the reward is 1. Therefore, the goal for discounted reward MDP is to reach the state 1 as soon as possible. And, the optimal policy is to take the third action in all states. Suppose we start in state $n-1$ with a uniform policy that takes all actions with equal probability. The policy gradient algorithm will try to improve this policy by taking a step in the gradient direction, where gradient

$$\nabla \rho(\pi) = \sum_s d^\pi(s) \sum_a Q^\pi(s,a) \nabla \pi(s,a)$$

Now $d^\pi(s)$ is the total discounted probability to be in state s at time $t = 1, 2, \dots$ according to the uniform policy π, and $Q^\pi(s,a)$ is the Q-value of this policy. Note that Q-value estimates will be 0 if we do not observe state 1 in any of the sample trajectories. The given policy takes two steps back and one step forward. By standard random walk analysis, the expected time to reach that goal state 1 from state n is exponential in n for such policy. Therefore, intuitively, to obtain non-zero estimates of gradient we need to observe on-policy sample trajectories of exponential length.

More formally, by standard random walk analysis, the probability to reach any state $s \le n/2$ is $(p/q)^{n/2} = \left(\frac{1/3}{2/3}\right)^{n/2} = (1/2)^{n/2}$. Therefore, $d^\pi(s) \le (1/2)^{n/2}$, i.e, exponentially small in n. Now, for any state $s \ge n/2$, $Q^\pi(s,a)$ is at most the probability to reach the state 1, i.e., $Q^\pi(s,a) \le (1/2)^{n/2}$. Therefore, each term in above summation is at most $\frac{1}{1-\gamma}(1/2)^{n/2}$, giving that the total magnitude of each component of the gradient is at most $\frac{1}{1-\gamma} n (1/2)^{n/2}$. Therefore, all components of the gradient must be estimated within an exponentially small error to get meaningful gradient estimates so that policy improvement can occur, requiring exponentially many samples.

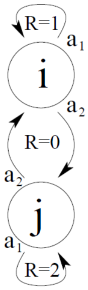

Figure 2: MDP for Example 2

1.2 Example 2

The second example illustrates a scenario with a small number of states (2 states), where a policy gradient based approach can still take exponential time to converge.

Consider Figure 1.2, there are 2 states, and 2 actions in each state. Action a_1 in each state loops back to the same state, with reward 1 in state i and reward 2 in state j. And, action a_2 transitions to the other state with reward 0. Let the goal is either to maximize infinite horizon average reward; or discounted infinite horizon reward with γ close to 1. In either case, the optimal policy is to play action a_2 in state i to transition to j; and action a_1 in state j. This policy has asymptotic average reward of 2.

Let the stationary distribution of initial policy π is

$$p(i) = 0.8, p(j) = 0.2.$$

An example of such a policy is $\pi(i, a_1) = 0.8, \pi(i, a_2) = 0.2$, $\pi(j, a_1) = 0.2, \pi(j, a_2) = 0.8$. Therefore, initially, the probability of being in state i is much higher than that of being in state j. On the other hand the optimal policy has stationary distribution with probability 1 on being in state j.

Consider policy improvement method using parametric policy of form $\pi(s, a) \propto e^{\theta^\top \phi_{s,a}}$, with $\phi_{s,a} = \mathbf{1}_{s,a}$. The optimal policy can be encoded by setting $\theta_{i,a_2} >> \theta_{i,a_1}$ and $\theta_{j,a_1} >> \theta_{j,a_2}$. Policy gradient is given by

$$\begin{aligned}
\nabla_\theta \rho^{\pi_\theta}(s_1) &= \mathbb{E}_{s\sim d^\pi, a\sim\pi(s)}[Q^\pi(s,a)\nabla_\theta \log(\pi(s,a))] \\
&= \mathbb{E}_{s\sim d^\pi, a\sim\pi(s)}[Q^\pi(s,a)(\phi_{s,a} - \sum_{a'} \pi(s,a')\phi_{s,a'})] \\
&= [d^\pi(s)\pi(s,a)Q^\pi(s,a)]_{s=i,j,a=a_1,a_2} - [d^\pi(s)\pi(s,a')\sum_a \pi(s,a)Q^\pi(s,a)]_{s=i,j,a'=a_1,a_2} \\
&= [d^\pi(s)\pi(s,a)Q^\pi(s,a)]_{s=i,j,a=a_1,a_2} - [d^\pi(s)\pi(s,a')V^\pi(s)]_{s=i,j,a'=a_1,a_2} \\
&= [d^\pi(s)\pi(s,a)(Q^\pi(s,a) - V^\pi(s))]_{s=i,j,a=a_1,a_2}
\end{aligned}$$

Update of $\theta_{s,a}$ by gradient ascent step:

$$\theta_{s,a} \leftarrow \theta_{s,a} + \alpha d^\pi(s)\pi_\theta(s,a)(Q^\pi(s,a) - V^\pi(s))$$

Note that the gradient depends on $d^\pi(s), s = i, j$, where $(1-\gamma)d^\pi(s)$ is roughly (for γ close to 1) the stationary distribution of the current policy π, and $Q^\pi(s,a)$. For state i, $Q^\pi(i, a_1)$ is higher that $Q^\pi(i, a_2)$ [1]. Therefore, θ_{i,a_1} increases compared to θ_{i,a_2}, which means the updated policy will favor looping on i even more rather than transitioning to j. For j, again θ_{j,a_1} increases more, but since the stationary probability $d^\pi(i) \approx \frac{1}{1-\gamma}p(i)$ is higher for state i than $d^\pi(j) \approx \frac{1}{1-\gamma}p(j)$, the state i gets an update of higher magnitude - moves more aggressively towards

[1] To see this, note that $Q^\pi(i, a_1) = 1 + \gamma V^\pi(i) = 1 + \gamma 0.8 + 0.8\gamma^2 V^\pi(i)) + \gamma^2 0.2 V^\pi(j)$, and $Q^\pi(i, a_2) = \gamma V^\pi(j) = \gamma^2 0.8 V^\pi(i) + 0.2 \times 2\gamma + 0.2\gamma^2 V^\pi(j) = Q^\pi(i, a_1) - 1 - 0.4\gamma$

the new θ. The next policy (updated θ) is likely to be even worse, i.e., the stationary probability is even higher for state i because of the increased probability of taking action a_1 in state i. Due to these reasons, initially the stationary probability of j decreases, and becomes exponentially small (see Figure 3.3 in Kakade and Langford [2002]), before it comes back to the correct policy in exponential time steps.

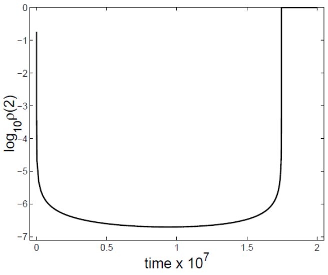

Figure 3: Stationary probability of state j: Figure 3.3 from Kakade and Langford [2002]

2 Provably efficient policy gradient methods

Above examples demonstrate that the improvement in policy in every step of the policy gradient can either be very small (example 1), or may even be negative, i.e., the policy may become worse (example 2, where initially, the new policy has increased probability of looping on state i). One underlying reason is that the gradient may not be an accurate measure of policy improvement (i.e., change in the gain of policy) when the policy is changed substantially.

The following lemma exactly quantifies the difference in gain for two policies $\tilde{\pi}$ and π. This lemma originally proven in Kakade and Langford [2002] is referred to as "performance difference lemma" in recent works (Agarwal et al. [2020]). This lemma lies at the heart of most analyses of policy gradient methods.

As earlier, for any policy π, let $d^\pi(s) = \sum_{t=1}^{\infty} \Pr(s_t = s|s_1)\gamma^{t-1}$. Note that $(1-\gamma)d^\pi$ defines a distribution over states.

Lemma 1 (Lemma 6.1 of Kakade and Langford [2002])**.** *For any two policies $\tilde{\pi}, \pi$,*

$$\rho(\tilde{\pi}) - \rho(\pi) = \sum_s d^{\tilde{\pi}}(s) \sum_a \tilde{\pi}(s,a) A^\pi(s,a) = \frac{1}{(1-\gamma)} \mathbb{E}_{s\sim(1-\gamma)d^{\tilde{\pi}}, a|s\sim\tilde{\pi}}[A^\pi(s,a)]$$

Here, $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$.

Proof.

$$
\begin{aligned}
\rho(\tilde{\pi}) = V^{\tilde{\pi}}(s_1) &= \mathbb{E}_{s_1,a_1,s_2,a_2,\ldots\sim\tilde{\pi}}[\sum_{t=1}^{\infty}\gamma^{t-1}R(s_t,a_t)|s_1] \\
&= \sum_{t=1}^{\infty}\gamma^{t-1}\mathbb{E}_{s_t,a_t\sim\tilde{\pi}}[R(s_t,a_t)+V^{\pi}(s_t)-V^{\pi}(s_t)|s_1] \\
&= \sum_{t=1}^{\infty}\gamma^{t-1}\mathbb{E}_{s_t,a_t\sim\tilde{\pi},s_{t+1}\sim P_{s_t,a_t}}[R(s_t,a_t)+\gamma V^{\pi}(s_{t+1})-V^{\pi}(s_t)|s_1]+V^{\pi}(s_1) \\
&= \sum_{t=1}^{\infty}\gamma^{t-1}\mathbb{E}_{s_t,a_t}[Q^{\pi}(s_t,a_t)-V^{\pi}(s_t)|s_1]+V^{\pi}(s_1) \\
&= \sum_{t=1}^{\infty}\gamma^{t-1}\mathbb{E}_{s_t,a_t}[A^{\pi}(s_t,a_t)|s_1]+V^{\pi}(s_1) \\
&= \sum_{s}d^{\tilde{\pi}}(s)\sum_{a}\tilde{\pi}(s,a)A^{\pi}(s,a)+V^{\pi}(s_1) \\
&= \sum_{s}d^{\tilde{\pi}}(s)\sum_{a}\tilde{\pi}(s,a)A^{\pi}(s,a)+\rho(\pi)
\end{aligned}
$$

□

A trouble with using the above lemma directly for algorithm design is that the state distribution in the expression on the right hand side is with respect to the new policy $\tilde{\pi}$, but the state distribution in data collected would be from the current policy π.

2.1 Conservative greedy algorithm

Kakade and Langford [2002] design a "conservative greedy" policy improvement method with guaranteed improvement in every step. Therefore, one can precisely quantify the number of steps required for this algorithm to converge.

This algorithm updates the policy in a lazy manner to ensure improvement with every update. Let π' be a new policy. (For example, this could be the policy obtained after gradient ascent from the old policy π. But, this could also be a policy that is very different from π). The conservative greedy algorithm makes the following lazy update to the policy:

$$\pi^{\text{new}}(s,a) := (1-\alpha)\pi(s,a)+\alpha\pi'(s,a) \tag{1}$$

Note that if α is small, π^{new} will be close to π even if π' is far from π. The parameter α provides a way to directly control how much you change π in each update. Kakade and Langford [2002] provide a lower bound on increase in gain for such an update for any π', π. Following, they suggest that the policy π' and step size α should be picked in such a way that this lower bound is maximized. The conservative greedy algorithm (with right choice of π' and step size α) is guaranteed to converge to a policy where no further local improvement can be made to the policy. Below we describe these results in detail.

Quantifying policy improvement

Following lemma lower bounds the gain from policy improvement.

Lemma 2 (Lemma 4.1 of Kakade and Langford [2002])**.**

$$\rho(\pi^{new})-\rho(\pi) \geq \alpha A_{\pi}(\pi') - \frac{2\alpha^2\gamma\epsilon}{(1-\gamma(1-\alpha))}$$

where $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$,

$$A_\pi(\pi') := \sum_s d^\pi(s) \sum_a \pi'(s,a) A^\pi(s,a) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim(1-\gamma)d^\pi, a\sim\pi'}[A^\pi(s,a)],$$

$$\epsilon := \frac{1}{(1-\gamma)} \left(\max_s \left| \sum_a \pi'(s,a) A^\pi(s,a) \right| \right).$$

Intuitively, $A_\pi(\pi')$*, measures to what extent advantage can increase if a different action (according to* π'*) was chosen in every visited state under* π*. And,* ϵ *is maximum of such advantage over different states.*

(Compare this to performance difference lemma. Note that the policy advantage uses the state distribution under current policy π instead of new policy $\tilde{\pi}$ or π'.)

Proof. We can use the policy gradient theorem to get an intuition for this. For a fixed π', there is one policy π_{new} for every value of $\alpha \in [0,1]$. Therefore, we can consider π^{new} as a parametric policy π^{new}_α parametrized by α; with $\pi^{new}_\alpha = \pi$ for $\alpha = 0$, and $\pi^{new}_\alpha = \pi'$ for $\alpha = 1$. Then, using policy gradient theorem, we get that gradient of $\rho(\pi^{new}_\alpha)$ at $\alpha = 0$ is given by

$$\begin{aligned}
\nabla_\alpha \rho(\pi^{new}_\alpha)\Big|_{\alpha=0} &= \sum_s d^{\pi^{new}_\alpha}(s) \sum_a \left(\nabla_\alpha \pi^{new}_\alpha(s,a)\right) A^{\pi^{new}_\alpha}(s,a)\Big|_{\alpha=0} \\
&= \sum_s d^{\pi^{new}_\alpha}(s) \sum_a (\pi'(s,a) - \pi(s,a)) A^{\pi^{new}_\alpha}(s,a)\Big|_{\alpha=0} \\
&= \sum_s d^\pi(s) \sum_a (\pi'(s,a) - \pi(s,a)) A^\pi(s,a) \\
&= \sum_s d^\pi(s) \sum_a \pi'(s,a) A^\pi(s,a) - \sum_s d^\pi(s) \pi(s,a) A^\pi(s,a) \\
&= \sum_s d^\pi(s) \sum_a \pi'(s,a) A^\pi(s,a) \\
&= A_\pi(\pi')
\end{aligned}$$

The second last step follows because $\sum_s d^\pi(s)\pi(s,a)A^\pi(s,a) = \sum_s d^\pi(s)\pi(s,a)(Q^\pi(s,a) - V^\pi(s)) = \sum_s d^\pi(s)(V^\pi(s) - V^\pi(s)) = 0$. In fact, by the same insight, $A_\pi(\pi^{new}) = \alpha A_\pi(\pi')$. Therefore, using Taylor expression, a lower bound on the improvement is given by

$$\rho(\pi^{new}_\alpha) - \rho(\pi) \geq \alpha A_\pi(\pi') - O(\alpha^2) = A_\pi(\pi^{new}) - O(\alpha^2)$$

To get a precise lower bound expression, we use the stronger performance difference lemma above (Lemma 1). This lemma shows that improvement in gain is given precisely by the following expression:

$$\rho(\pi^{new}) - \rho(\pi) = \sum_s d^{\pi^{\text{new}}}(s) \sum_a \pi^{\text{new}}(s,a) A^\pi(s,a)$$

Compare this to

$$\alpha A_\pi(\pi') = A_\pi(\pi^{new}) := \sum_s d^\pi(s) \sum_a \pi^{new}(s,a) A^\pi(s,a)$$

The first expression above uses state distribution under π^{new} instead of π. To get a precise lower bound we need to bound the error due to this measure mismatch.

To compare the two, a coupling argument is used.In any given state s, π^{new} picks actions according to π' with probability α and according to π with probability $1-\alpha$. Now, for any fixed time t, let η_t be the number of steps before time t where π^{new} did not take action according to π, i.e., η_t is the number of mismatches in the actions suggested by π^{new} and π. Then, conditional on event $\eta_t = 0$, the distribution of states before time t is same for trajectories generated from π^{new} and π. More precisely,

$$\Pr_{\tau\sim\pi^{\text{new}}}(s_t = s|\eta_t = 0) = \Pr_{\tau\sim\pi}(s_t = s)$$

where random variable $\tau = (s_1, s_2, \ldots, s_t, \ldots)$ denotes a trajectory . Further, the probability that there was some mismatch before t is given by $p_t := \Pr(\eta_t > 0) = 1 - \Pr(\eta_t = 0) = 1 - (1-\alpha)^{t-1}$. Therefore,

$$\begin{aligned}
\rho(\pi^{new}) - \rho(\pi) &= \sum_s d^{\pi^{\text{new}}}(s) \sum_a \pi^{\text{new}}(s,a) A^\pi(s,a) \\
&= \mathbb{E}_{\tau=(s_1,s_2,\ldots,)\sim\pi^{\text{new}}}[\sum_t \gamma^{t-1} \sum_a \pi^{\text{new}}(s_t,a) A^\pi(s_t,a)] \\
&= \mathbb{E}_{\tau=(s_1,s_2,\ldots,)\sim\pi^{\text{new}}}[\sum_t \gamma^{t-1} \sum_a \alpha\pi'(s_t,a) A^\pi(s_t,a)] \\
&= \alpha \sum_t (1-p_t)\gamma^{t-1} \mathbb{E}_{\tau=(s_1,s_2,\ldots,s_t)\sim\pi^{\text{new}}}[\sum_a \pi'(s_t,a) A^\pi(s_t,a)|\eta_t = 0] \\
&\quad +\alpha \sum_t p_t \gamma^{t-1} \mathbb{E}_{\tau=(s_1,s_2,\ldots,)\sim\pi^{\text{new}}}[\sum_a \pi'(s_t,a) A^\pi(s_t,a)|\eta_t > 0] \\
&= \alpha \sum_t (1-p_t)\gamma^{t-1} \mathbb{E}_{(s_1,s_2,\ldots,s_t)\sim\pi}[\sum_a \pi'(s_t,a) A^\pi(s_t,a)] \\
&\quad +\alpha \sum_t p_t \gamma^{t-1} \mathbb{E}_{\tau=(s_1,s_2,\ldots,)\sim\pi^{\text{new}}}[\sum_a \pi'(s_t,a) A^\pi(s_t,a)|\eta_t > 0] \\
&= \alpha A_\pi(\pi') - \alpha \sum_t p_t \gamma^{t-1} \mathbb{E}_{(s_1,s_2,\ldots,s_t)\sim\pi}[\sum_a \pi'(s_t,a) A^\pi(s_t,a)] \\
&\quad +\alpha \sum_t \gamma^{t-1} p_t \mathbb{E}_{\tau=(s_1,s_2,\ldots,)\sim\pi^{\text{new}}}[\sum_a \pi'(s_t,a) A^\pi(s_t,a)|\eta_t > 0] \\
&\geq \alpha A_\pi(\pi') - 2\alpha \sum_t (1-(1-\alpha)^{t-1})\gamma^{t-1} \left(\max_s \left| \sum_a \pi'(s,a) A^\pi(s,a) \right| \right) \\
&= \alpha A_\pi(\pi') - 2\alpha\epsilon(1-\gamma) \left(\frac{1}{1-\gamma} - \frac{1}{1-(1-\alpha)\gamma} \right) \\
&= \alpha A_\pi(\pi') - 2\alpha\epsilon \frac{\alpha\gamma}{(1-(1-\alpha)\gamma)}
\end{aligned}$$

where $\epsilon = \frac{1}{1-\gamma} \left(\max_s |\sum_a \pi'(s,a) A^\pi(s,a)|\right)$.

□

Algorithm design: selecting π' and α

Lemma 2 provides a lower bound on the improvement for a given α and π'

Choice of α. For $\alpha = 1$ (greedy update), the improvement is lower bounded as:

$$\rho(\pi^{new}) - \rho(\pi) \geq A_\pi(\pi') - 2\gamma\epsilon$$

However, the second quantity can be larger than the first in above, as ϵ is an upper bound on $A_\pi(\pi')$, and therefore, the above improvement may be negative.

However, we can select a step size to ensure positive improvement as long as $A_\pi(\pi') > 0$. Let R be an upper bound on rewards, so that $A_\pi(\pi') = \sum_s d^\pi(s)\pi'(s,a)A^\pi(s,a) \leq \frac{R}{(1-\gamma)^2}, \epsilon \leq \frac{R}{(1-\gamma)^2}$. Then, setting

$$\alpha = \frac{A_\pi(\pi')(1-\gamma)^3}{4R}, \tag{2}$$

and substituting in Lemma 2, we get

$$\rho(\pi^{new}) - \rho(\pi) \geq \alpha A_\pi(\pi') - \frac{2\alpha^2\epsilon}{(1-\gamma)} \geq \frac{A_\pi(\pi')^2(1-\gamma)^3}{8R} \tag{3}$$

Remark: Note that since the 'distribution' $d^\pi(s)$ in the definition of $A_\pi(\pi')$ is not normalized to 1, and instead sums $1/(1-\gamma)$, and $A^\pi(s,a) \leq \frac{1}{1-\gamma}$, $A_\pi(\pi')$ in the above expression is upper bounded by $R/(1-\gamma)^2$. So, the above expression for the lower bound on improvement is of order $\frac{R}{1-\gamma}$, i.e., of the same order as the value function.

Choice of π'. From above quantification, to ensure maximum improvement, π' should maximize $A_\pi(\pi')$. Intuitively, $A_\pi(\pi')$ measures to what extent advantage can increase if a different action (according to π') was chosen in every visited state under π. Clearly,

$$\max_{\pi'} A_\pi(\pi') = \sum_s d^\pi(s) \max_a A^\pi(s,a)$$

Therefore, the policy that maximizes policy advantage is given by

$$\pi'(s) = \arg\max_a A^\pi(s,a).$$

However, to use this policy the advantage $A^\pi(s,a)$ needs to be estimated. The following algorithm allows approximate estimation of $A^\pi(s,a)$:

Algorithm. Initialize π.

1. Set $\hat{A} = \sum_s d^\pi(s) \max_a \hat{A}^\pi(s,a)$, where $\hat{A}^\pi(s,a)$ are estimates of $A^\pi(s,a)$ for every s, a and such that

   $$(1-\gamma)^2 |\hat{A} - \max_{\pi'} A_\pi(\pi')| \leq \frac{\delta}{3}$$

   Here $A_\pi(\pi') = \sum_s d^\pi(s) \max_a A^\pi(s,a)$. This can be done for example by estimating $A^\pi(s,a)$ as function approximation $\hat{A}^\pi = f_\omega(s,a)$ where parameter ω is set through sample estimation with loss function (normalized to have values in $[-R, +R]$):

   $$(1-\gamma)^2 \sum_s d^\pi(s) \max_a |A^\pi(s,a) - f_\omega(s,a)|$$

   Since this is an expected error over state distribution under the current policy π, this loss can be approximated using trajectory samples from the current policy (with sufficient exploration over actions to be able to estimate $A^\pi(s,a)$ for $a \neq \pi(s)$). Roughly $\frac{R^2}{\delta^2} \log \frac{R^2}{\delta^2}$ samples are required to ensure a δ with probability $1-\delta$.

2. If $(1-\gamma)^2 \hat{A} < \frac{2\delta}{3}$, STOP.

3. Otherwise, update policy:

   $$\pi \leftarrow (1-\alpha)\pi + \alpha\pi'$$

   where

   $$\pi'(s) := \arg\max_a f_\omega(s,a).$$

   $$\alpha = \left((1-\gamma)^2 \hat{A} - \frac{\delta}{3}\right) \frac{(1-\gamma)}{4R}$$

4. Go back to Step 1.

In above procedure, in every iteration $(1-\gamma)^2 A_\pi(\pi') \geq \hat{A}(1-\gamma)^2 - \frac{\delta}{3} \geq \frac{2\delta}{3} - \frac{\delta}{3} = \frac{\delta}{3}$, therefore, from (3), the increase in gain is at least

$$\left(\hat{A}(1-\gamma)^2 - \frac{\delta}{3}\right)^2 \frac{1}{8R(1-\gamma)} \geq \frac{\delta^2}{72R(1-\gamma)}$$

Since the total improvement to be made is at the most maximum value of gain, i.e., $R/(1-\gamma)$, the procedure terminates in at most $\frac{R}{1-\gamma} \frac{72R(1-\gamma)}{\delta^2} = \frac{72R^2}{\delta^2}$ steps. That is, we have the following result.

Lemma 3. *Above conservative greedy algorithm terminates in at most $\frac{72R^2}{\delta^2}$ steps to find a policy π such that*

$$(1-\gamma)\max_{\pi'} A_\pi(\pi') \leq \delta$$

3 How good is the policy? local improvement vs. optimality

As demonstrated in the last section, the conservative greedy algorithm (with right choice of π' and step size α) is guaranteed to terminate. It terminates at the policy π such that $(1-\gamma)\max_{\pi'} A_\pi(\pi') \leq \delta$. This can be interpreted as the condition that there is no (or very little) advantage increase on changing the policy under the state distribution of the current policy. In other words, it cannot be improved locally. But, how does this policy compare to the "optimal policy", which may have completely different state distribution.

The following theorem shows that the gap can be large if the stationary distribution over states for the chosen policy is very different from the stationary distribution over states for the optimal policy. This can happen if there isn't enough exploration over states. The following theorem also provides a way to ensure exploration. It states that one could start from a different starting state distribution (e.g. uniform) than the target starting state distribution, and then the gap depends only on how the stationary distribution of optimal policy differs from the uniform distribution.

Theorem 4 (Theorem 6.2 of Kakade and Langford [2002])**.** *Let $d^{\pi,\mu}$ denotes the (non-normalized) stationary distribution over states for policy π when starting state distribution is μ. Also, let $\rho(\pi;\mu)$ denote the gain of policy π when starting state distribution is given by μ.*

Suppose we have a policy π, with $(1-\gamma)^2\max_{\pi'} A_\pi(\pi') \leq \delta$, where $A_\pi(\pi') = \sum_s d^{\pi,\mu}(s)\sum_a \pi'(s,a)A^\pi(s,a)$. Then, for any policy π^* and starting state distribution μ^*,

$$\begin{aligned}\rho(\pi^*;\mu^*) - \rho(\pi;\mu^*) &\leq \frac{\delta}{(1-\gamma)^2}\left\|\frac{d^{\pi^*,\mu^*}}{d^{\pi,\mu}}\right\|_\infty \\ &\leq \frac{\delta}{(1-\gamma)^2}\left\|\frac{(1-\gamma)d^{\pi^*,\mu^*}}{\mu}\right\|_\infty\end{aligned}$$

Proof.

$$\begin{aligned}\rho(\pi^*;\mu^*) - \rho(\pi;\mu^*) &= \sum_s d^{\pi^*,\mu^*}(s)\sum_a \pi^*(s,a)A^\pi(s,a) \\ &= \sum_s \frac{d^{\pi^*,\mu^*}(s)}{d^{\pi,\mu}(s)} d^{\pi,\mu}(s)\sum_a \pi^*(s,a)A^\pi(s,a) \\ &\leq \left\|\frac{d^{\pi^*,\mu^*}}{d^{\pi,\mu}}\right\|_\infty \sum_s d^{\pi,\mu}(s)\max_a A^\pi(s,a) \\ &\leq \left\|\frac{d^{\pi^*,\mu^*}}{d^{\pi,\mu}}\right\|_\infty \frac{\delta}{(1-\gamma)^2} \\ &\leq \left\|\frac{d^{\pi^*,\mu^*}}{\mu}\right\|_\infty \frac{\delta}{(1-\gamma)^2}\end{aligned}$$

The last step follows from the observation that $d^{\pi,\mu}(s) \geq \Pr(s_1 = s;\pi,\mu) = \mu(s)$. □

Therefore, if we can choose the starting state distribution to be a uniform distribution, we can bound the gap from the optimal policy by $\frac{n\delta}{(1-\gamma)^3}$, where n is the number of states. (Some applications may not have the flexibility of choosing the starting state distribution).

4 Trust Region Policy Optimization [Schulman et al., 2015]

The TRPO algorithm from Schulman et al. [2015] can be explained as a simple (and useful!) extension of the above ideas. The paper considers ideas to get rid of *mixed* policies of form $\pi^{\text{new}} = (1-\alpha)\pi + \alpha\pi'$ used in the conservative greedy algorithm described above. Note that every iteration of the conservative greedy algorithm stated adds a new policy to this mixture, making the collection of policies to maintain potentially quite large and inconvenient. Instead can we just maintain policy parameter θ, and simply update the policy parameter θ to obtain a new policy? Can the new policy still have properties similar to π^{new} in terms of policy improvement and closeness to policy π?

Schulman et al. [2015] provide the following lower bound on the improvement in gain when policy π is updated to arbitrary policy $\tilde{\pi}$, which is very similar to the quantification in Lemma 2 (even has essentially the same proof). But, this way of quantifying the gap allows finding a new policy $\tilde{\pi}$ close to π through parameter search in a small 'trust' region around the parameter for π, instead of restricting to mixed policies.

Lemma 5 (Theorem 1 of Schulman et al. [2015])**.** *For any two policies $\tilde{\pi}$ and π,*

$$\begin{aligned}\rho(\tilde{\pi}) - \rho(\pi) &\geq A_\pi(\tilde{\pi}) - \frac{2\epsilon\gamma}{(1-\gamma)}\, D_{TV}^{\max}(\pi, \tilde{\pi})^2 \\ &\geq A_\pi(\tilde{\pi}) - \frac{4\epsilon\gamma}{(1-\gamma)}\, D_{KL}^{\max}(\pi, \tilde{\pi})\end{aligned}$$

with $\epsilon = \frac{1}{(1-\gamma)} \max_s \max_a |A^\pi(s,a)|$; and $D_{TV}^{\max}(\pi, \tilde{\pi}) = \max_s D_{TV}(\pi(s,\cdot)||\tilde{\pi}(s,\cdot))$ is the maximum total variation distance between the two policies, and $D_{KL}^{\max}(\pi, \tilde{\pi}) = \max_s D_{KL}(\pi(s,\cdot)||\tilde{\pi}(s,\cdot))$ is the maximum KL divergence, over all states.

Proof. The proof is almost exactly the same as the proof of Lemma 2. Consider again the proof of Lemma 2. We observe that all the steps of the proof of Lemma 2 follow for arbitrary policies π and $\pi^{\text{new}} := \tilde{\pi}$, by setting α such that in any state s, $\Pr_{a\sim\pi(s),\tilde{a}\sim\tilde{\pi}(s)}(a \neq \tilde{a}|s) \leq \alpha$.

Below we provide the proof details. As before, we couple the trajectories of π and $\tilde{\pi}$. Let p_t be the probability that $\tilde{\pi}$ suggests a different action than π in some time step before t. We denote this event as $\eta_t > 0$. Then $p_t := \Pr(\eta_t > 0) \leq 1 - (1-\alpha)^{t-1}$. Now,

$$\begin{aligned}\rho(\tilde{\pi}) - \rho(\pi) &= \sum_s d^{\tilde{\pi}}(s) \sum_a \tilde{\pi}(s,a) A^\pi(s,a) \\ &= \mathbb{E}_{\tau=(s_1,s_2,\ldots,)\sim\tilde{\pi}}[\sum_t \gamma^{t-1} \sum_a \tilde{\pi}(s_t,a) A^\pi(s_t,a)] \\ &= \sum_t (1-p_t)\gamma^{t-1} \mathbb{E}_{\tau=(s_1,s_2,\ldots,s_t)\sim\tilde{\pi}}[\sum_a \tilde{\pi}(s_t,a) A^\pi(s_t,a)|\eta_t = 0] \\ &\quad + \sum_t p_t \gamma^{t-1} \mathbb{E}_{\tau=(s_1,s_2,\ldots,)\sim\tilde{\pi}}[\sum_a \tilde{\pi}(s_t,a) A^\pi(s_t,a)|\eta_t > 0] \\ &= \sum_t (1-p_t)\gamma^{t-1} \mathbb{E}_{(s_1,s_2,\ldots,s_t)\sim\pi}[\sum_a \tilde{\pi}(s_t,a) A^\pi(s_t,a)] \\ &\quad + \sum_t p_t \gamma^{t-1} \mathbb{E}_{\tau=(s_1,s_2,\ldots,)\sim\tilde{\pi}}[\sum_a \tilde{\pi}(s_t,a) A^\pi(s_t,a)|\eta_t > 0] \\ &= A_\pi(\tilde{\pi}) - \sum_t p_t \gamma^{t-1} \mathbb{E}_{(s_1,s_2,\ldots,s_t)\sim\pi}[\sum_a \tilde{\pi}(s_t,a) A^\pi(s_t,a)] \\ &\quad + \sum_t \gamma^{t-1} p_t \mathbb{E}_{\tau=(s_1,s_2,\ldots,)\sim\tilde{\pi}}[\sum_a \tilde{\pi}(s_t,a) A^\pi(s_t,a)|\eta_t > 0] \\ &\geq A_\pi(\tilde{\pi}) - 2\sum_t p_t \gamma^{t-1} \left(\max_s \sum_a \tilde{\pi}(s,a) A^\pi(s,a)\right) \\ &\geq A_\pi(\tilde{\pi}) - 2\tilde{\epsilon}(1-\gamma) \sum_t (1-(1-\alpha)^{t-1})\gamma^{t-1} \\ &= A_\pi(\tilde{\pi}) - 2\tilde{\epsilon}\left(\frac{1}{1-\gamma} - \frac{1}{1-(1-\alpha)\gamma}\right) \\ &= A_\pi(\tilde{\pi}) - 2\tilde{\epsilon}\frac{\alpha\gamma}{(1-(1-\alpha)\gamma)}\end{aligned}$$

where $\tilde{\epsilon} := \frac{1}{1-\gamma}\left(\max_s \sum_a \tilde{\pi}(s,a)A^\pi(s,a)\right)$. Now, using $\Pr_{a\sim\pi,\tilde{a}\sim\tilde{\pi}}(a \neq \tilde{a}|s) \leq \alpha$, we get that

$$\tilde{\epsilon} = \frac{1}{1-\gamma}\left(\max_s \sum_a \tilde{\pi}(s,a)A^\pi(s,a) - \pi(s,a)A^\pi(s,a)\right) \leq \frac{1}{1-\gamma}\left(\max_s \max_{a,\tilde{a}} \alpha|A^\pi(s,a) - A^\pi(s,\tilde{a})|\right) \leq 2\alpha\epsilon$$

where $\epsilon = \frac{1}{1-\gamma}\max_{s,a}|A^\pi(s,a)|$. Substituting,

$$\rho(\tilde{\pi}) - \rho(\pi) \quad \geq \quad A_\pi(\tilde{\pi}) - \frac{2\alpha^2\epsilon\gamma}{(1-\gamma)}$$

Now, setting $\alpha = D_{TV}^{\max}(\pi,\tilde{\pi})$, we have $\Pr(a \neq \tilde{a}|s) \leq \alpha$. This gives the stated result. □

Note that the above lemma is very similar in form to Lemma 2 which considered $\tilde{\pi}$ as mixed policy with α being the mixing parameter, and provided:

$$\rho(\tilde{\pi}) - \rho(\pi) \quad \geq \quad A_\pi(\tilde{\pi}) - \frac{2\epsilon\gamma}{(1-\gamma)}\alpha^2$$

(For mixed policy $\tilde{\pi} = (1-\alpha)\pi + \alpha\pi'$, $A_\pi(\tilde{\pi}) = \alpha A_\pi(\pi')$, and $\epsilon = \frac{1}{1-\gamma}\left(\max_s \sum_a \pi'(s,a)A^\pi(s,a)\right)$). The main contribution of Lemma 5 is to get rid of mixing through α and provide bounds in terms of total variation distance for arbitrary updates.

Given the above result, the paper proposes the following way to update the policy parameter. A natural strategy to find a good policy is to search over all policies with small total variation distance or KL-divergence from the policy π and find the one with maximum $A_\pi(\tilde{\pi})$. Let π is parameterized by θ. And, $A_\theta(\tilde{\theta})$ denote $A_{\pi_\theta}(\tilde{\pi}_{\tilde{\theta}})$. Then, TRPO algorithm chooses next policy by solving

$$\begin{aligned} \max_{\tilde{\theta}} \quad & A_\theta(\tilde{\theta}) \\ \text{s.t.} \quad & D_{KL}^{\max}(\pi_\theta||\pi_{\tilde{\theta}}) \leq \eta \end{aligned}$$

The region of parameters with $D_{KL}^{\max}(\pi_\theta,\pi_{\tilde{\theta}}) \leq \eta$ is referred to as the trust region.

References

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift, 2020.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, pages 267–274, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1-55860-873-7. URL `http://dl.acm.org/citation.cfm?id=645531.656005`.

John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1889–1897. JMLR.org, 2015. URL `http://dl.acm.org/citation.cfm?id=3045118.3045319`.