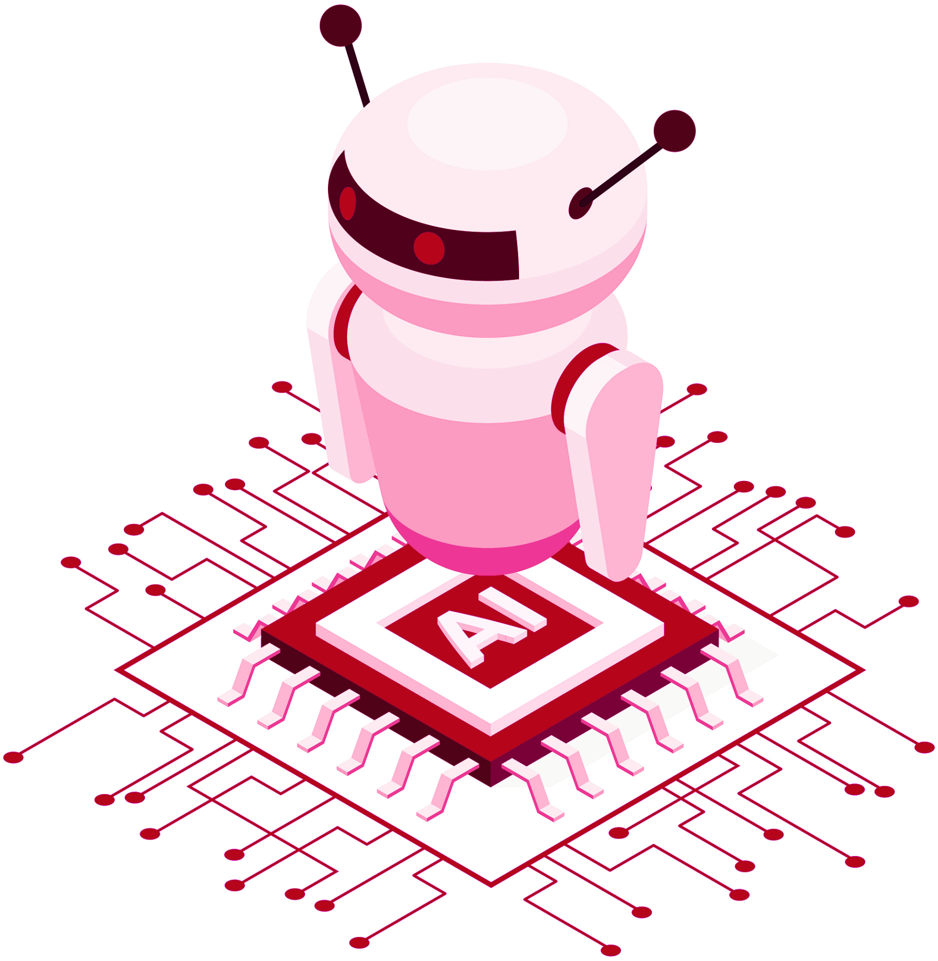

# Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Homework 8:

## Policy-Based Theory

By:

[Asemaneh Nafe]

[400105285]

Spring 2025

# Contents

# Grading

The grading will be based on the following criteria, with a total of 100 points:

| **Task** | **Points** |
|---|---|
| Policy Gradient - Part (a) | 20 |
| Policy Gradient - Part (b) | 10 |
| Trust Region Policy Optimization - Part (a) | 10 |
| Trust Region Policy Optimization - Part (b) | 5 |
| Trust Region Policy Optimization - Part (c) | 10 |
| Trust Region Policy Optimization - Part (d) | 20 |
| Trust Region Policy Optimization - Part (e) | 20 |
| Trust Region Policy Optimization - Part (f) | 5 |
| Bonus: Writing your report in Latex | 5 |

# 1 Policy Gradient Theorem

In this question, we will prove the policy gradient theorem and provide a set of sufficient conditions that allow us to use function approximations as a critic for the $Q$-value function so that the policy gradient using our function approximation remains exact.

## 1.1 Notations

Consider a normal finite MDP with bounded rewards. $P(s'|s,a)$ represents the transition model, which corresponds to the probability of transitioning from state $s$ to $s'$ due to action $a$. Also, the reward model is represented by $r : \mathcal{S} \times \mathcal{A} \to [0,1]$ where $r(s,a)$ is the immediate reward associated with taking action $a$ in state $s$. Parameter $\gamma \in [0,1)$ corresponds to the discount factor, and $s_0$ indicates the starting state of our MDP.

A parametrized policy $\pi_\theta$ induces a distribution over trajectories $\tau = (s_t, a_t, r_t)|_{t=0}^{\infty}$ where $s_0$ is the starting state, and for all subsequent timesteps $t$, $a_t \sim \pi(.|s_t)$, $s_{t+1} \sim P(.|s_t, a_t)$. The state value function and the state-action value ($Q$-value) functions are defined as follows by the Bellman operator:

$$V^{\pi_\theta}(s) = \mathbb{E}_{a \sim \pi_\theta(.|s)}[Q^{\pi_\theta}(s,a)]$$
$$Q^{\pi_\theta}(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(.|s,a)}[V^{\pi_\theta}(s')]$$

We also define the discounted state visitation distribution $d^{\pi}_{s_0}$ of a policy $\pi$ as:

$$d^{\pi}_{s_0}(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t Pr^{\pi}(s_t = s|s_0), \tag{1}$$

where $Pr^{\pi}(s_t = s|s_0)$ is the state visitation probability that $s_t = s$, after we execute $\pi$ starting at state $s_0$.

## 1.2 Proving the Policy Gradient Theorem

The objective function of our RL problem is defined as $J(\theta) = V^{\pi_\theta}(s_0)$. The policy gradient method uses the gradient ascent algorithm to optimize $\theta$. This can be done by the direct differentiation of the objective function.

a) Prove the following identity, which is known as the Policy Gradient Theorem:

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s \sim d^{\pi_\theta}_{s_0}} \mathbb{E}_{a \sim \pi_\theta(.|s)}[\nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a)] \tag{2}$$

we know that:

$$J(\theta) = V^{\pi_\theta}(s_0)$$

From the definition of the state-value function:

$$V^{\pi_\theta}(s) = \sum_a \pi_\theta(a|s) Q^{\pi_\theta}(s,a)$$

$$\nabla_\theta V^{\pi_\theta}(s) = \sum_a \nabla_\theta \pi_\theta(a|s) Q^{\pi_\theta}(s,a) + \sum_a \pi_\theta(a|s) \nabla_\theta Q^{\pi_\theta}(s,a)$$

$$Q^{\pi_\theta}(s,a) = r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^{\pi_\theta}(s') \Rightarrow \nabla_\theta Q^{\pi_\theta}(s,a) = \gamma \sum_{s'} P(s'|s,a) \nabla_\theta V^{\pi_\theta}(s')$$

Substitute into our gradient expression:

$$\nabla_\theta V^{\pi_\theta}(s) = \sum_a \nabla_\theta \pi_\theta(a|s) Q^{\pi_\theta}(s,a) + \gamma \sum_a \pi_\theta(a|s) \sum_{s'} P(s'|s,a) \nabla_\theta V^{\pi_\theta}(s')$$

This can be rewritten as:

$$\nabla_\theta V^{\pi_\theta}(s) = \sum_a \nabla_\theta \pi_\theta(a|s) Q^{\pi_\theta}(s,a) + \gamma \sum_{s'} P^{\pi_\theta}(s'|s) \nabla_\theta V^{\pi_\theta}(s')$$

This recursive structure allows us to unroll the gradient through time. Thus:

$$\nabla_\theta V^{\pi_\theta}(s_0) = \sum_{t=0}^{\infty} \gamma^t \sum_{s_t} P^{\pi_\theta}(s_t|s_0) \sum_{a_t} \nabla_\theta \pi_\theta(a_t|s_t) Q^{\pi_\theta}(s_t,a_t)$$

Applying the log-derivative identity:

$$\nabla_\theta \pi_\theta(a|s) = \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s)$$

Then the expression becomes:

$$\nabla_\theta V^{\pi_\theta}(s_0) = \sum_{t=0}^{\infty} \gamma^t \sum_{s_t} P^{\pi_\theta}(s_t|s_0) \sum_{a_t} \pi_\theta(a_t|s_t) \nabla_\theta \log \pi_\theta(a_t|s_t) Q^{\pi_\theta}(s_t,a_t)$$

Now define the $\gamma$-discounted state visitation distribution as:

$$d_{s_0}^{\pi_\theta}(s) := (1-\gamma) \sum_{t=0}^{\infty} \gamma^t P^{\pi_\theta}(s_t = s|s_0)$$

This leads us to:

$$\sum_{t=0}^{\infty} \gamma^t P^{\pi_\theta}(s_t = s|s_0) = \frac{d_{s_0}^{\pi_\theta}(s)}{1-\gamma}$$

Substituting back into the gradient of $J(\theta)$:

$$\nabla_\theta J(\theta) = \nabla_\theta V^{\pi_\theta}(s_0) = \frac{1}{1-\gamma} \sum_s d_{s_0}^{\pi_\theta}(s) \sum_a \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a)$$

Finally, expressing this using expectation notation:

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \left[ \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left[ \nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a) \right] \right]$$

## 1.3 Compatible Function Approximation Theorem

Now, consider the case in which $Q^{\pi_\theta}$ is approximated by a learned function approximator. If the approximation is sufficiently good, we might hope to use it in place of $Q^{\pi_\theta}$ in equation 2. If we use the function approximator $Q_\phi(s,a)$, the convergence of our method is not necessarily maintained due to the fact that our gradient will not be exact anymore. The following theorem provides sufficient conditions for our function approximator so that our gradient using the approximator remains exact.

**Theorem 1.1** *(Compatible Function Approximation). If the following two conditions are satisfied for any function approximator with parameter $\phi$:*

1. *Critic gradient is compatible with the Actor score function, i.e.,*

$$\nabla_\phi Q_\phi(s,a) = \nabla_\theta \log \pi_\theta(a|s)$$

2. *Critic parameters $\phi$ minimize the following mean-squared error[1]:*

$$\epsilon = \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(.|s)} [(Q^{\pi_\theta}(s,a) - Q_\phi(s,a))^2]$$

*Then, the policy gradient using critic $Q_\phi(s,a)$ is exact, i.e.,*

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} [\nabla_\theta \log \pi_\theta(a|s) Q_\phi(s,a)]$$

b) Prove theorem 1.1.

We aim to prove that under the two conditions of the theorem, the policy gradient using the function approximator $Q_\phi(s,a)$ is exact, i.e.,

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left[ \nabla_\theta \log \pi_\theta(a|s) Q_\phi(s,a) \right]$$

We begin by recalling the standard form of the policy gradient:

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left[ \nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a) \right]$$

Let us define a function approximator $Q_\phi(s,a)$ to approximate $Q^{\pi_\theta}(s,a)$. Define the mean squared error between the true and approximated Q-functions as:

$$\epsilon(\phi) = \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left[ \left( Q^{\pi_\theta}(s,a) - Q_\phi(s,a) \right)^2 \right]$$

We assume that $\phi$ minimizes this error, and that the minimum is unique.

[1] Assume that the mean-squared error has only one critical point which corresponds to its minimum.

Now, take the gradient of $\epsilon(\phi)$ with respect to $\phi$ and set it to zero at the minimum:

$$\nabla_\phi \epsilon(\phi) = -2\,\mathbb{E}_{s \sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[ \left( Q^{\pi_\theta}(s,a) - Q_\phi(s,a) \right) \nabla_\phi Q_\phi(s,a) \right] = 0$$

Using the **compatibility condition**, which states that

$$\nabla_\phi Q_\phi(s,a) = \nabla_\theta \log \pi_\theta(a|s)$$

we substitute this into the expression:

$$\mathbb{E}_{s \sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[ \left( Q^{\pi_\theta}(s,a) - Q_\phi(s,a) \right) \nabla_\theta \log \pi_\theta(a|s) \right] = 0$$

so

$$\mathbb{E}_{s \sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[ \left( Q^{\pi_\theta}(s,a) \right) \nabla_\theta \log \pi_\theta(a|s) \right] = \mathbb{E}_{s \sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[ \left( Q_\phi(s,a) \right) \nabla_\theta \log \pi_\theta(a|s) \right]$$

So now consider the original policy gradient:

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s,a} \left[ \nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a) \right]$$

Hence,

$$\nabla_\theta J(\theta) * (1-\gamma) = \mathbb{E}_{s,a} \left[ \nabla_\theta \log \pi_\theta(a|s) Q_\phi(s,a) \right]$$
$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s,a} \left[ \nabla_\theta \log \pi_\theta(a|s) Q_\phi(s,a) \right]$$

which completes the proof.

# 2 Trust Region Policy Optimization

In this question, we will dive deep into the mathematical theories behind the TRPO algorithm. As a roadmap, we first prove that minimizing a certain surrogate objective function guarantees policy improvement with non-trivial step sizes. Then, we make a series of approximations to the theoretically justified algorithm, yielding a practical algorithm, which has been called trust region policy optimization (TRPO).

## 2.1 Notations and Preliminaries

Let $\pi$ denote a stochastic policy and let $\eta(\pi)$ denote its expected discounted reward:

$$\eta(\pi) = \mathbb{E}_{s_0,a_0,\ldots}[\sum_{t=0}^{\infty} \gamma^t r(s_t)]$$

where

$$s_0 \sim \rho_0(s_0), a_t \sim \pi(a_t|s_t), s_{t+1} \sim P(s_{t+1}|s_t, a_t).$$

Also, we will use the following standard definitions of the state-action value function $Q_\pi$, the value function $V_\pi$, and the advantage function $A_\pi$:

$$Q_\pi(s_t, a_t) = \mathbb{E}_{s_{t+1},a_{t+1},\ldots}[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l})]$$
$$V_\pi(s_t) = \mathbb{E}_{a_t,s_{t+1},\ldots}[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l})]$$
$$A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s)$$

a) Prove the following identity:

$$\eta(\pi') = \eta(\pi) + \mathbb{E}_{s_0,a_0,\ldots\sim\pi'}[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t)] \tag{3}$$

proof:

$$\begin{aligned}
\eta(\pi') - \eta(\pi) &= \eta(\pi') - E_{s_0 \sim p(s_0)}[v^\pi(s_0)] && (4)\\
&= \eta(\pi') - E_{\tau \sim p_{\theta'}}[v^\pi(s_0)] && (5)\\
&= \eta(\pi') - E_{\tau \sim p_{\theta'}}[\sum_{t=0}^{\infty} \gamma^t v^\pi(s_t) - \sum_{t=1}^{\infty} \gamma^t v^\pi(s_t)] && (6)\\
&= E_{\tau \sim p_{\theta'}}[\sum_{t=0}^{\infty} \gamma^t r(s_t)] - E_{\tau \sim p_{\theta'}}[\sum_{t=0}^{\infty} \gamma^t (v^\pi(s_t) + \gamma v^\pi(s_t + 1))] && (7)\\
&= E_{\tau \sim p_{\theta'}}[\sum_{t=0}^{\infty} \gamma^t (r(s_t) + \gamma v^\pi(s_t + 1) - v^\pi(s_t)] && (8)\\
&= E_{\tau \sim p_{\theta'}}[\sum_{t=0}^{\infty} \gamma^t (Q^\pi(s_t, a_t) - v^\pi(s_t)] && (9)
\end{aligned}$$

so:

$$\eta(\pi') = \eta(\pi) + \mathbb{E}_{s_0, a_0, \ldots \sim \pi'}[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t)]$$

other way to proof(in the article): fig [1] First note that $A_\pi(s,a) = \mathbb{E}_{s' \sim P(s'|s,a)}\left[r(s) + \gamma V_\pi(s') - V_\pi(s)\right]$. Therefore,

$$\begin{aligned}
\mathbb{E}_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t A_\pi\left(s_t, a_t\right)\right] &= \mathbb{E}_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t (r(s_t) + \gamma V_\pi(s_{t+1}) - V_\pi(s_t))\right] && (10)\\
&= \mathbb{E}_{\tau|\tilde{\pi}}\left[-V_\pi(s_0) + \sum_{t=0}^{\infty} \gamma^t r(s_t)\right] && (11)\\
&= -\mathbb{E}_{s_0}\left[V_\pi(s_0)\right] + \mathbb{E}_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t)\right] && (12)\\
&= -\eta(\pi) + \eta(\tilde{\pi}) && (13)
\end{aligned}$$

Rearranging, the result follows.

Equation 3 basically shows that the difference between the expected total rewards of any two policies $\pi'$ and $\pi$ depends on the advantage function of policy $\pi$ if the trajectory is sampled by running $\pi'$. We will use this equation to derive an optimization scheme further to maximize the expected total reward using the advantage function of policy $\pi$ to obtain policy $\pi'$.

Let $\rho_\pi$ be the unnormalized discounted visitation frequencies:

$$\rho_\pi(s) = P(s_0 = s) + \gamma P(s_1 = s) + \gamma^2 P(s_2 = s) + \ldots$$

b) Prove the following identity:

$$\eta(\pi') = \eta(\pi) + \sum_s \rho_{\pi'}(s) \sum_a \pi'(a|s) A_\pi(s,a) \tag{14}$$

proof:

$$
\begin{aligned}
\eta(\tilde{\pi}) &= \eta(\pi) + \sum_{t=0}^{\infty} \sum_{s} P(s_t = s \mid \tilde{\pi}) \sum_{a} \tilde{\pi}(a \mid s) \gamma^t A_\pi(s,a) && (15) \\
&= \eta(\pi) + \sum_{s} \sum_{t=0}^{\infty} \gamma^t P(s_t = s \mid \tilde{\pi}) \sum_{a} \tilde{\pi}(a \mid s) A_\pi(s,a) && (16) \\
&= \eta(\pi) + \sum_{s} \rho_{\tilde{\pi}}(s) \sum_{a} \tilde{\pi}(a \mid s) A_\pi(s,a). && (17)
\end{aligned}
$$

Equation 14 can be used as an optimization objective in reinforcement learning. Note that this equation has been considered difficult to optimize directly due to the complex dependency of $\rho_{\pi'}(s)$ on $\pi'$. Instead, the following local approximation of $\eta$ has been introduced for optimization:

$$
L_\pi(\pi') = \eta(\pi) + \sum_{s} \rho_{\pi}(s) \sum_{a} \pi'(a|s) A_\pi(s,a) \tag{18}
$$

Note that $L_\pi$ uses the visitation frequency $\rho_\pi$ rather than $\rho_{\pi'}$, ignoring changes in state visitation density due to changes in the policy. In the next section, we will derive an algorithm to guarantee a monotonic improvement in our policy using equation 18 as our objective function, showing that equation 18 is good enough in our case.

## 2.2 Monotonic Improvement Guarantee for General Stochastic Policies

In this section, we build the theoretical foundations to consider the policy optimization problem, assuming that the policy can be evaluated at all states. The ultimate goal of this section is to prove the following theorem:

**Theorem 2.1** *Let $\pi, \pi'$ be two stochastic policies. Then, the following bound holds:*

$$
\eta(\pi') \geq L_\pi(\pi') - \frac{4\epsilon\gamma}{(1-\gamma)^2} D_{KL}^{\max}(\pi, \pi')
$$
$$
\textit{where } \epsilon = \max_{s,a} |A_\pi(s,a)|
$$

During this section, we use the following definitions and inequality for the total variation and KL divergence:

$$
\begin{aligned}
&D_{TV}(p||q) = \frac{1}{2} \sum_i |p_i - q_i| \\
&D_{TV}^{\max}(\pi, \pi') = \max_s D_{TV}(\pi(.|s)||\pi'(.|s)) \\
&D_{KL}^{\max}(\pi, \pi') = \max_s D_{KL}(\pi(.|s)||\pi'(.|s)) \\
&D_{TV}(p||q)^2 \leq D_{KL}(p||q)
\end{aligned}
$$

We will prove theorem 2.1 step by step, and you are required to complete the proof as indicated below. To begin the proof, we denote trajectories by $\tau$ and define $\bar{A}(s)$ as follows:

$$\bar{A}(s) = \mathbb{E}_{a\sim\pi'(.|s)}[A_\pi(s,a)]$$

Then we can rewrite equations 14 and 18 as follows:

$$\eta(\pi') = \eta(\pi) + \mathbb{E}_{\tau\sim\pi'}[\sum_{t=0}^{\infty}\gamma^t \bar{A}(s_t)] \tag{19}$$

$$L_\pi(\pi') = \eta(\pi) + \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty}\gamma^t \bar{A}(s_t)] \tag{20}$$

The only difference in these two equations is whether the states are sampled using $\pi$ or $\pi'$. To bound the difference between $\eta(\pi')$ and $L_\pi(\pi')$, we first need to introduce a measure of how much $\pi$ and $\pi'$ agree. Specifically, we'll couple the policies so that they define a joint distribution over pairs of actions. We use the following definition of $\alpha$-coupled policy pairs:

**Definition 2.2** *$(\pi, \pi')$ is an $\alpha$-coupled policy pair if it defines a joint distribution $(a, a')|s$ such that $P(a \neq a'|s) \leq \alpha$ for all $s$. $\pi$ and $\pi'$ will denote the marginal distributions of $a$ and $a'$, respectively.*

c) Prove the following lemma:

**Lemma 2.3** *Given that $\pi, \pi'$ are $\alpha$-coupled policies, for all $s$,*

$$|\bar{A}(s)| \leq 2\alpha \max_{s,a} |A_\pi(s,a)|$$

$$\bar{A}(s) = \mathbb{E}_{\tilde{a}\sim\tilde{\pi}}\left[A_\pi(s,\tilde{a})\right] = \mathbb{E}_{(a,\tilde{a})\sim(\pi,\tilde{\pi})}\left[A_\pi(s,\tilde{a}) - A_\pi(s,a)\right] \text{ since } \mathbb{E}_{a\sim\pi}\left[A_\pi(s,a)\right] = 0 \tag{21}$$

$$= P(a \neq \tilde{a} \mid s)\mathbb{E}_{(a,\tilde{a})\sim(\pi,\tilde{\pi})|a\neq\tilde{a}}\left[A_\pi(s,\tilde{a}) - A_\pi(s,a)\right] \tag{22}$$

$$|\bar{A}(s)| \leq \alpha \cdot 2\max_{s,a}|A_\pi(s,a)| \tag{23}$$

d) Prove the following lemma:

**Lemma 2.4** *Let $(\pi, \pi')$ be an $\alpha$-coupled policy pair. Then:*

$$|\mathbb{E}_{s_t\sim\pi'}[\bar{A}(s_t)] - \mathbb{E}_{s_t\sim\pi}[\bar{A}(s_t)]| \leq 4\alpha(1-(1-\alpha)^t)\max_{s,a}|A_\pi(s,a)|$$

Proof. Given the coupled policy pair $(\pi, \tilde{\pi})$, we can also obtain a coupling over the trajectory distributions produced by $\pi$ and $\tilde{\pi}$, respectively. Namely, we have pairs of trajectories $\tau, \tilde{\tau}$, where $\tau$ is obtained by taking actions from $\pi$, and $\tilde{\tau}$ is obtained by taking actions from $\tilde{\pi}$, where the same random seed is used to generate both trajectories. We will consider the advantage of $\tilde{\pi}$ over $\pi$ at timestep $t$, and decompose this expectation based on whether $\pi$ agrees with $\tilde{\pi}$ at all timesteps $i < t$.

Let $n_t$ denote the number of times that $a_i \neq \tilde{a}_i$ for $i < t$, i.e., the number of times that $\pi$ and $\tilde{\pi}$ disagree before timestep $t$.

$$\mathbb{E}_{s_t\sim\tilde{\pi}}\left[\bar{A}(s_t)\right] = P(n_t = 0)\mathbb{E}_{s_t\sim\tilde{\pi}|n_t=0}\left[\bar{A}(s_t)\right] + P(n_t > 0)\mathbb{E}_{s_t\sim\tilde{\pi}|n_t>0}\left[\bar{A}(s_t)\right] \tag{24}$$

The expectation decomposes similarly for actions are sampled using $\pi$:

$$\mathbb{E}_{s_t\sim\pi}\left[\bar{A}(s_t)\right] = P(n_t = 0)\mathbb{E}_{s_t\sim\pi|n_t=0}\left[\bar{A}(s_t)\right] + P(n_t > 0)\mathbb{E}_{s_t\sim\pi|n_t>0}\left[\bar{A}(s_t)\right] \tag{25}$$

Note that the $n_t = 0$ terms are equal:

$$\mathbb{E}_{s_t\sim\pi|n_t=0}\left[\bar{A}(s_t)\right] = \mathbb{E}_{s_t\sim\tilde{\pi}|n_t=0}\left[\bar{A}(s_t)\right], \tag{26}$$

because $n_t = 0$ indicates that $\pi$ and $\tilde{\pi}$ agreed on all timesteps less than $t$. Subtracting Equations (24) and (25), we get

$$\mathbb{E}_{s_t\sim\tilde{\pi}}\left[\bar{A}(s_t)\right] - \mathbb{E}_{s_t\sim\pi}\left[\bar{A}(s_t)\right] = P(n_t > 0)\left(\mathbb{E}_{s_t\sim\tilde{\pi}|n_t>0}\left[\bar{A}(s_t)\right] - \mathbb{E}_{s_t\sim\pi|n_t>0}\left[\bar{A}(s_t)\right]\right) \tag{27}$$

By definition of $\alpha, P(\pi, \tilde{\pi}$ agree at timestep $i) \geq 1 - \alpha$, so $P(n_t = 0) \geq (1-\alpha)^t$, and

$$P(n_t > 0) \leq 1 - (1-\alpha)^t \tag{28}$$

Next, note that

$$\begin{aligned} |\mathbb{E}_{s_t\sim\tilde{\pi}|n_t>0}\left[\bar{A}(s_t)\right] - \mathbb{E}_{s_t\sim\pi|n_t>0}\left[\bar{A}(s_t)\right]| &\leq |\mathbb{E}_{s_t\sim\tilde{\pi}|n_t>0}\left[\bar{A}(s_t)\right]| + |\mathbb{E}_{s_t\sim\pi|n_t>0}\left[\bar{A}(s_t)\right]| && (29)\\ &\leq 4\alpha \max_{s,a} |A_\pi(s,a)| && (30) \end{aligned}$$

Where the second inequality follows from Lemma 2.3.

Plugging Equation (28) and Equation (30) into Equation (27), we get

$$|\mathbb{E}_{s_t\sim\tilde{\pi}}\left[\bar{A}(s_t)\right] - \mathbb{E}_{s_t\sim\pi}\left[\bar{A}(s_t)\right]| \leq 4\alpha(1-(1-\alpha)^t)\max_{s,a}|A_\pi(s,a)| \tag{31}$$

e) Prove the following lemma:

**Lemma 2.5** *Let $(\pi, \pi')$ be an $\alpha$-coupled policy pair. Then:*

$$|\eta(\pi') - L_\pi(\pi')| \leq \frac{4\alpha^2\gamma\epsilon}{(1-\gamma)^2}$$

The preceding Lemma bounds the difference in expected advantage at each timestep $t$. We can sum over time to bound the difference between $\eta(\tilde{\pi})$ and $L_\pi(\tilde{\pi})$. Subtracting Equation (19) and Equation (20) , and defining $\epsilon = \max_{s,a}|A_\pi(s,a)|$,

$$\begin{aligned} |\eta(\tilde{\pi}) - L_\pi(\tilde{\pi})| &= \sum_{t=0}^{\infty} \gamma^t |\mathbb{E}_{\tau\sim\tilde{\pi}}\left[\bar{A}(s_t)\right] - \mathbb{E}_{\tau\sim\pi}\left[\bar{A}(s_t)\right]| && (32)\\ &\leq \sum_{t=0}^{\infty} \gamma^t \cdot 4\epsilon\alpha(1-(1-\alpha)^t) && (33)\\ &= 4\epsilon\alpha\left(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\alpha)}\right) && (34)\\ &= \frac{4\alpha^2\gamma\epsilon}{(1-\gamma)(1-\gamma(1-\alpha))} && (35)\\ &\leq \frac{4\alpha^2\gamma\epsilon}{(1-\gamma)^2} && (36) \end{aligned}$$

f) Prove theorem 2.1. Hint: Use the fact that if we have two policies $\pi$ and $\pi'$ such that $D_{TV}^{\max}(\pi, \pi') \leq \alpha$, then we can define an $\alpha$-couples policy pair $(\pi, \pi')$ with appropriate marginals.[2]

For each state $s$, define the total variation distance between the action distributions of policies $\pi$ and $\pi'$ as:

$$\alpha(s) := D_{\mathsf{TV}}(\pi(\cdot \mid s) \,\|\, \pi'(\cdot \mid s)).$$

By **Pinsker's inequality**, we know that:

$$D_{\mathsf{TV}}(p \,\|\, q) \leq \sqrt{\frac{1}{2} D_{\mathsf{KL}}(p \,\|\, q)},$$

as Dkl is always posetive;

$$D_{\mathsf{TV}}(p \,\|\, q) \leq \sqrt{D_{\mathsf{KL}}(p \,\|\, q)},$$

which implies:

$$\alpha(s) \leq \sqrt{D_{\mathsf{KL}}(\pi(\cdot \mid s) \,\|\, \pi'(\cdot \mid s))}.$$

Now define:

$$\alpha_{\max} := \max_s \alpha(s).$$

Using the inequality above, we can bound $\alpha_{\max}$ as:

$$\alpha_{\max} \leq \max_s \sqrt{D_{\mathsf{KL}}(\pi(\cdot \mid s) \,\|\, \pi'(\cdot \mid s))} \leq \sqrt{D_{\mathsf{KL}}^{\max}(\pi, \pi')},$$

where:

$$D_{\mathsf{KL}}^{\max}(\pi, \pi') := \max_s D_{\mathsf{KL}}(\pi(\cdot \mid s) \,\|\, \pi'(\cdot \mid s)).$$

From the hint, given such an upper bound on total variation, there exists an $\alpha_{\max}$-coupled policy pair $(\pi, \pi')$ where for each state $s$, a joint distribution over actions $(a, a')$ satisfies:

$$\mathbb{P}(a \neq a' \mid s) \leq \alpha_{\max}.$$

Now, we invoke the assumed Lemma 2.5, which states that under an $\alpha$-coupled policy pair:

$$|\eta(\pi') - L_\pi(\pi')| \leq \frac{4\alpha^2 \gamma \epsilon}{(1-\gamma)^2},$$

where $\epsilon = \max_{s,a} |A_\pi(s, a)|$.

from the bound on $\alpha_{\max}$, we get:

$$\alpha_{\max}^2 \leq D_{\mathsf{KL}}^{\max}(\pi, \pi'),$$

so:

$$|\eta(\pi') - L_\pi(\pi')| \leq \frac{4\gamma\epsilon}{(1-\gamma)^2} \cdot D_{\mathsf{KL}}^{\max}(\pi, \pi')$$

This gives us:

$$\eta(\pi') \geq L_\pi(\pi') - \frac{4\gamma\epsilon}{(1-\gamma)^2} D_{\mathsf{KL}}^{\max}(\pi, \pi').$$

[2] There is no need to prove this hint!

Note that the inequality in theorem 2.1 becomes an equality in $\pi' = \pi$. Thus, the following optimization problem guarantees a non-decreasing expected return $\eta$:

$$\begin{aligned}
&\pi_{i+1} = \arg\max_{\pi} L_{\pi_i}(\pi) - C D_{KL}^{\max}(\pi_i, \pi) \\
&\text{where } C = \frac{4\epsilon\gamma}{(1-\gamma)^2} \\
&\text{and } L_{\pi_i}(\pi) = \eta(\pi_i) + \sum_s \rho_{\pi_i}(s) \sum_a \pi(a|s) A_{\pi_i}(s,a)
\end{aligned}$$

In practice, if we use the penalty coefficient $C$ as recommended by the theory above, the step sizes would be very small. One way to take larger steps in a robust way is to use a constraint on the KL divergence between the two policies as a trust region:

$$\begin{aligned}
&\pi_{i+1} = \arg\max_{\pi} L_{\pi_i}(\pi) \\
&\text{subject to } D_{KL}^{\max}(\pi_i, \pi) \leq \delta
\end{aligned}$$

This problem imposes a constraint that the KL divergence is bounded at every point in the state space. While it is motivated by the theory, this problem is impractical to solve due to the large number of constraints. Instead, we can use a heuristic approximation by considering the average KL divergence. The following optimization problem has been proposed as the TRPO algorithm:

$$\begin{aligned}
&\pi_{i+1} = \arg\max_{\pi} L_{\pi_i}(\pi) \\
&\text{subject to } \mathbb{E}_{s\sim\rho}[D_{KL}(\pi_i(.|s)||\pi(.|s))] \leq \delta
\end{aligned}$$