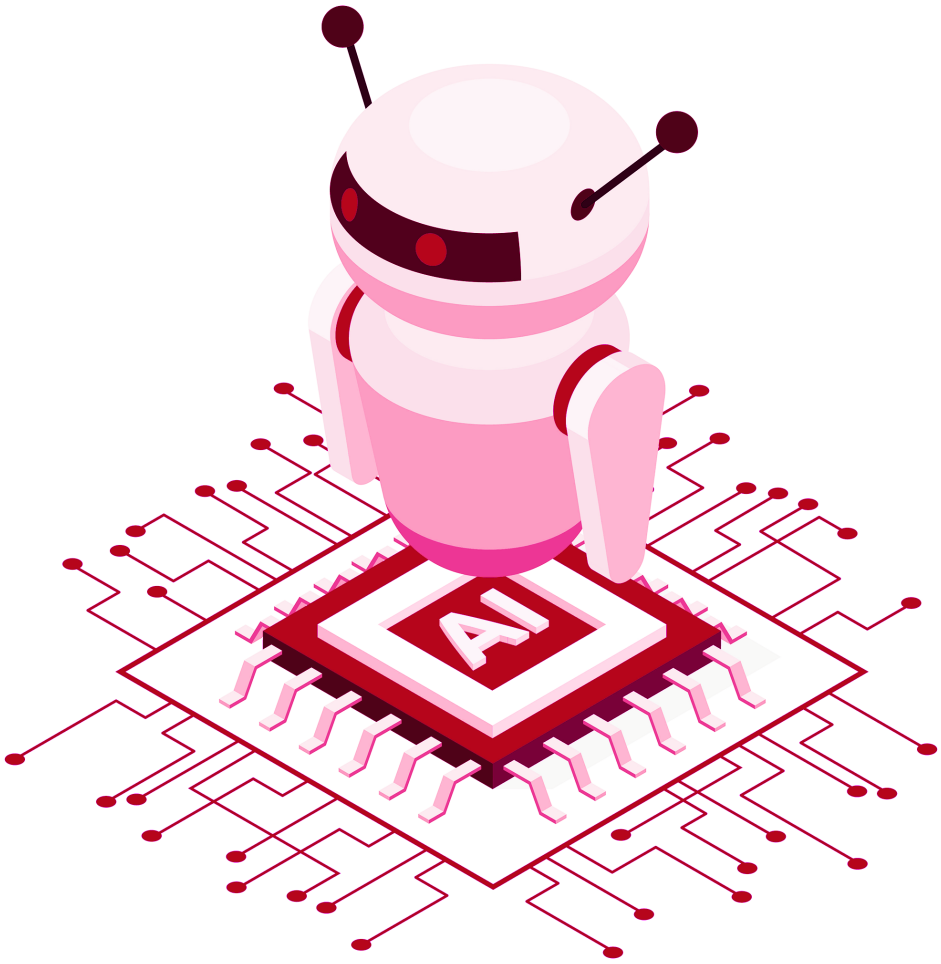

Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Homework 8:

Policy-Based Theory

By:

Payam Taebi

400104867

RIML

Deep RL Course

Spring 2025

Contents

1 Policy Gradient Theorem **1**

1.1 Notations 1

1.2 Proving the Policy Gradient Theorem 1

1.3 Compatible Function Approximation Theorem 3

2 Trust Region Policy Optimization **6**

2.1 Notations and Preliminaries 6

2.2 Monotonic Improvement Guarantee for General Stochastic Policies 10

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 π_θ 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 $d_{s_0}^{\pi}$ of a policy π as:

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

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

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

Solution to Question (a)

Definitions and preliminaries

Let

$$J(\theta) \;=\; V^{\pi_\theta}(s_0) \;=\; \mathbb{E}_{\tau\sim\pi_\theta}\Big[\sum_{t=0}^{\infty}\gamma^t\, r(s_t,a_t) \;\Big|\; s_0\Big].$$

We recall two standard identities that we shall invoke repeatedly:

1. **Log-derivative (likelihood–ratio) trick**

$$\nabla_\theta \pi_\theta(a|s) \;=\; \pi_\theta(a|s)\,\nabla_\theta \log \pi_\theta(a|s). \tag{LR}$$

2. **Discounted state–visitation distribution**

$$d_{s_0}^{\pi_\theta}(s) \;=\; (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \; \overset{\pi_\theta}{\Pr}\big[s_t = s \mid s_0\big]. \tag{D}$$

Throughout, rewards are bounded and $\gamma \in [0,1)$, ensuring all expectations exist and are finite.

Step-by-step derivation

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}_{\tau\sim\pi_\theta}\big[r(s_t,a_t) \mid s_0\big] \\
&\overset{(1)}{=} \sum_{t=0}^{\infty} \gamma^t \sum_{s} \overset{\pi_\theta}{\Pr}(s_t = s \mid s_0) \sum_{a} \nabla_\theta \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a),
\end{aligned} \tag{3}$$

where step (1) substitutes the definition of the Q-function, $Q^{\pi_\theta}(s,a) = \mathbb{E}\big[\sum_{k=0}^{\infty}\gamma^k r_{t+k} \;\big|\; s_t = s, a_t = a\big]$.

Intuition. Equation (3) rewrites the gradient as a sum over every time-step, every state visited at that time, and every action chosen in that state.

Collecting time-steps via $d_{s_0}^{\pi_\theta}$. Using definition (D),

$$(1-\gamma)\sum_{t=0}^{\infty}\gamma^t \overset{\pi_\theta}{\Pr}(s_t = s \mid s_0) \;=\; d_{s_0}^{\pi_\theta}(s).$$

Hence

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma}\sum_{s} d_{s_0}^{\pi_\theta}(s) \sum_{a} \nabla_\theta \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a). \tag{4}$$

Applying the log-derivative trick. Insert identity (LR) inside (4):

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

Therefore

$$\begin{aligned}
\nabla_\theta J(\theta) &= \frac{1}{1-\gamma}\sum_{s} d_{s_0}^{\pi_\theta}(s) \sum_{a} \pi_\theta(a|s)\,\nabla_\theta \log \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a) \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}}\Big[\,\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\big[\nabla_\theta \log \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a)\big]\Big].
\end{aligned} \tag{5}$$

Alternate derivation. One may begin from the trajectory log-probability $\log p_\theta(\tau) = \sum_t \log \pi_\theta(a_t|s_t)$, differentiate inside the expectation for $J(\theta)$, and apply the law of total expectation to arrive at the same result. The state-distribution route above is chosen because it highlights the role of $d_{s_0}^{\pi_\theta}(s)$ and easily generalises to continuing-tasks and off-policy variants.

Conclusion

Combining (5) we obtain the **Policy Gradient Theorem**:

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

[Proof recap] Starting from the definition of $J(\theta)$, we introduced the discounted visitation distribution to aggregate over time-steps, invoked the log-derivative trick to isolate $\nabla_\theta \log \pi_\theta(a|s)$, and rearranged terms into nested expectations, arriving precisely at (2). All manipulations are justified because rewards are bounded and $\gamma < 1$, guaranteeing uniform convergence of the series and interchange of summation with differentiation.

Practical note. The theorem shows that any unbiased estimator of $Q^{\pi_\theta}(s,a)$ may be plugged into the inner expectation. Later parts of this problem set will examine conditions under which a learned *critic* $Q_w(s,a)$ preserves exactness of the gradient.

1.3 Compatible Function Approximation Theorem

Now, consider the case in which Q^{π_θ} is approximated by a learned function approximator. If the approximation is sufficiently good, we might hope to use it in place of Q^{π_θ} 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 ϕ:*

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 ϕ minimize the following mean-squared error*[1]*:*

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

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

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

b) Prove theorem 1.1.

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

Solution to Question (b)

[Compatible Function Approximation] Assume

1. **Compatibility (score–function match)**

$$\boxed{\nabla_\phi Q_\phi(s,a) \;=\; \nabla_\theta \log \pi_\theta(a \mid s)} \tag{C1}$$

2. **Critic optimality (MSE minimiser)**

$$\boxed{\phi \;=\; \arg\min_{\tilde{\phi}} \; \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \big[\big(Q^{\pi_\theta}(s,a) - Q_{\tilde{\phi}}(s,a) \big)^2 \big]} \tag{C2}$$

Then

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

Let the *residual* between the true and approximate action–value be

$$\delta(s,a) \;=\; Q^{\pi_\theta}(s,a) - Q_\phi(s,a).$$

Step 1: Optimality of ϕ implies an orthogonality condition. Define the mean-squared error functional

$$\varepsilon(\phi) \;=\; \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \big[\delta(s,a)^2 \big].$$

Because (C2) declares that ϕ is the *unique* critical point minimising $\varepsilon(\phi)$, we must have

$$\nabla_\phi \varepsilon(\phi) = 0.$$

Compute the gradient:

$$\begin{aligned} \nabla_\phi \varepsilon(\phi) &= \mathbb{E}_{s,a} \Big[2\, \delta(s,a)\, \nabla_\phi \big(-Q_\phi(s,a) \big) \Big] \\ &= -2\, \mathbb{E}_{s,a} \big[\delta(s,a) \; \nabla_\phi Q_\phi(s,a) \big]. \end{aligned} \tag{6}$$

Setting (6) to zero yields the *orthogonality condition*

$$\mathbb{E}_{s,a} \big[\delta(s,a) \; \nabla_\phi Q_\phi(s,a) \big] \;=\; 0. \tag{O}$$

Step 2: Invoke compatibility to link actor and critic. By (C1), $\nabla_\phi Q_\phi(s,a) = \nabla_\theta \log \pi_\theta(a \mid s)$. Substituting into (O),

$$\mathbb{E}_{s,a} \big[\delta(s,a) \; \nabla_\theta \log \pi_\theta(a \mid s) \big] \;=\; 0. \tag{O$'$}$$

Intuition. Condition (O′) states that, in expectation, the Bellman-error residual is *uncorrelated* with the policy's score function.

Step 3: Show the approximate gradient equals the true gradient. Recall the policy-gradient theorem (proved in part (a)):

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma}\,\mathbb{E}_{s,a}\big[\nabla_\theta \log \pi_\theta(a \mid s)\; Q^{\pi_\theta}(s,a)\big].$$

Add and subtract $Q_\phi(s,a)$ inside the expectation:

$$\begin{aligned}\nabla_\theta J(\theta) &= \tfrac{1}{1-\gamma}\mathbb{E}_{s,a}\big[\nabla_\theta \log \pi_\theta(a \mid s)\; \big(Q_\phi(s,a) + \delta(s,a)\big)\big] \\ &= \tfrac{1}{1-\gamma}\mathbb{E}_{s,a}\big[\nabla_\theta \log \pi_\theta(a \mid s)\; Q_\phi(s,a)\big] \;+\; \tfrac{1}{1-\gamma}\underbrace{\mathbb{E}_{s,a}\big[\nabla_\theta \log \pi_\theta(a \mid s)\; \delta(s,a)\big]}_{=\,0 \text{ by } (\mathrm{O}')}. \end{aligned} \tag{7}$$

The second term vanishes; hence

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

which is precisely the statement of the theorem.

Take-away intuition. The two compatibility conditions guarantee that the critic's *parameter-space gradient* spans the same subspace as the actor's score function. When the critic is trained to minimise mean-squared Bellman error, its residual becomes orthogonal to that subspace, so inserting Q_ϕ in place of Q^{π_θ} leaves the ascent direction unchanged—even though $Q_\phi \neq Q^{\pi_\theta}$ in general.

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 π 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_π, the value function V_π, and the advantage function A_π:

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

a) Prove the following identity:

$$\eta(\pi') = \eta(\pi) + \mathbb{E}_{s_0, a_0, \ldots \sim \pi'}[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t)] \tag{8}$$

Equation 8 basically shows that the difference between the expected total rewards of any two policies π' and π depends on the advantage function of policy π if the trajectory is sampled by running π'. We will use this equation to derive an optimization scheme further to maximize the expected total reward using the advantage function of policy π to obtain policy π'.

Goal. Let π be a reference policy and π' any other policy. Show that their expected discounted returns satisfy

$$\boxed{\eta(\pi') \;=\; \eta(\pi) \;+\; \mathbb{E}_{\tau\sim\pi'}\Big[\sum_{t=0}^{\infty}\gamma^{t}\,A_{\pi}(s_t,a_t)\Big]} \tag{1}$$

where $\tau = (s_0, a_0, s_1, a_1, \dots)$ is a (possibly infinite) trajectory generated by first sampling $s_0 \sim \rho_0$ and then following π'.

1. Recap of key definitions

- **Return:** $\eta(\pi) \;=\; \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^{t} r(s_t)\Big]$.
- **Value function:** $V_{\pi}(s) \;=\; \mathbb{E}_{\substack{a\sim\pi(\,\cdot\,|s)\\ \tau'\sim\pi}}\Big[\sum_{l=0}^{\infty}\gamma^{l} r(s_l)\Big]$, where τ' is the future trajectory that starts in s.
- **Action–value:** $Q_{\pi}(s,a) \;=\; \mathbb{E}_{\substack{s'\sim P(\,\cdot\,|s,a)\\ \tau'\sim\pi}}\Big[r(s) + \sum_{l=1}^{\infty}\gamma^{l} r(s_l)\Big]$.
- **Advantage:** $A_{\pi}(s,a) = Q_{\pi}(s,a) - V_{\pi}(s)$.

  Equivalently, one can express the immediate reward in terms of the advantage:

$$r(s) + \gamma\,\mathbb{E}_{s'\sim P}[\,V_{\pi}(s')\,] \;=\; A_{\pi}(s,a) + V_{\pi}(s). \tag{2}$$

2. Unroll the advantage along a trajectory

For a *fixed* trajectory τ (generated by *any* policy), sum (2) with a discount factor γ^t at each time step:

$$\sum_{t=0}^{\infty}\gamma^{t} A_{\pi}(s_t,a_t) = \sum_{t=0}^{\infty}\gamma^{t}\Big[r(s_t) + \gamma\,\mathbb{E}_{s_{t+1}\sim P}[V_{\pi}(s_{t+1})] - V_{\pi}(s_t)\Big]. \tag{3}$$

Telescoping trick. Shift the index in the middle term ($k = t+1$) and rewrite:

$$\sum_{t=0}^{\infty}\gamma^{t+1} V_{\pi}(s_{t+1}) = \sum_{k=1}^{\infty}\gamma^{k} V_{\pi}(s_k).$$

Hence the *second* and *third* sums in (3) cancel everywhere except at $t = 0$:

$$\sum_{t=0}^{\infty}\gamma^{t} A_{\pi}(s_t,a_t) = \underbrace{\sum_{t=0}^{\infty}\gamma^{t} r(s_t)}_{\text{cumulative reward}} - \underbrace{V_{\pi}(s_0)}_{\text{single leftover term}}. \tag{4}$$

3. Take expectations under the new policy π'

Let the trajectory τ now be distributed according to π'. Taking expectation on both sides of (4) gives

$$\begin{aligned}\mathbb{E}_{\tau\sim\pi'}\Big[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\Big] &= \mathop{\mathbb{E}}_{\tau\sim\pi'}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t)\Big] - \mathop{\mathbb{E}}_{s_0\sim\rho_0}\big[V_\pi(s_0)\big]\\ &= \eta(\pi') - \eta(\pi).\end{aligned}$$

Re-arranging proves (1).

4. Why is this identity useful?

1. It expresses *policy improvement* purely in terms of the *advantage* of an *older* policy π, evaluated on trajectories from the *newer* policy π'. Thus we need not know $A_{\pi'}$ to compare π' to π.
2. In practice one substitutes the intractable expectation $\mathbb{E}_{\tau\sim\pi'}$ with Monte-Carlo estimates or importance sampling, obtaining a *surrogate objective* that is maximised instead of $\eta(\pi')$ itself.
3. The identity is exact—no approximations so far. Later, TRPO introduces a *trust-region* constraint (bounding the KL divergence between π and π') that guarantees the surrogate objective is a *lower bound* on the true return, enabling monotonic improvement even with finite steps.

5. Intuition in one sentence

If the new policy π' spends more (discounted) time taking actions that are **better-than-average** under the old policy π (i.e. have positive A_π), it must outperform π; otherwise it cannot do worse because the negative contributions are exactly offset in the telescoping sum.

Let ρ_π be the unnormalized discounted visitation frequencies:

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

b) Prove the following identity:

$$\eta(\pi') = \eta(\pi) + \sum_s \rho_{\pi'}(s) \sum_a \pi'(a|s) A_\pi(s,a) \tag{9}$$

Equation 9 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 π'. Instead, the following local approximation of η has been introduced for optimization:

$$L_\pi(\pi') = \eta(\pi) + \sum_s \rho_{\pi}(s) \sum_a \pi'(a|s) A_\pi(s,a) \tag{10}$$

Note that L_π uses the visitation frequency ρ_π 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 10 as our objective function, showing that equation 10 is good enough in our case.

Goal. Show that for any two policies π (reference) and π' (candidate)

$$\eta(\pi') = \eta(\pi) + \sum_s \rho_{\pi'}(s) \sum_a \pi'(a \mid s)\, A_\pi(s,a), \tag{1}$$

where $\rho_{\pi'}(s) = \sum_{t=0}^{\infty} \gamma^t \Pr_{\pi'}(s_t = s)$ is the *discounted visitation frequency* of s under π' and A_π is the advantage w. r. t. the baseline policy π.

1. Start from the trajectory identity

From part (a) we already have

$$\eta(\pi') = \eta(\pi) + \mathbb{E}_{\tau\sim\pi'}\Big[\sum_{t=0}^{\infty} \gamma^t\, A_\pi(s_t,a_t)\Big]. \tag{2}$$

Thus we only need to rewrite the expectation on the right-hand side in terms of state–action frequencies.

2. Pull out the expectation step by step

Let $f(s,a) \coloneqq A_\pi(s,a)$ be *any* bounded function (initially the advantage; the derivation will hold for any f). Compute

$$\mathbb{E}_{\tau\sim\pi'}\Big[\sum_{t=0}^{\infty} \gamma^t f(s_t,a_t)\Big] = \sum_{t=0}^{\infty} \gamma^t \underbrace{\mathbb{E}_{\tau\sim\pi'}[f(s_t,a_t)]}_{\text{expected value at time } t}.$$

Condition on s_t. By the law of total probability,

$$\mathbb{E}_{\tau\sim\pi'}[f(s_t,a_t)] = \sum_s \Pr_{\pi'}(s_t = s) \sum_a \pi'(a \mid s)\, f(s,a).$$

Insert into the series and regroup.

$$\begin{aligned}\mathbb{E}_{\tau\sim\pi'}\Big[\sum_{t=0}^{\infty} \gamma^t f(s_t,a_t)\Big] &= \sum_{t=0}^{\infty} \gamma^t \sum_s \Pr_{\pi'}(s_t = s) \sum_a \pi'(a \mid s)\, f(s,a) \\ &= \sum_s \underbrace{\Big[\sum_{t=0}^{\infty} \gamma^t \Pr_{\pi'}(s_t = s)\Big]}_{=\,\rho_{\pi'}(s)} \sum_a \pi'(a \mid s)\, f(s,a).\end{aligned}$$

Apply $f = A_\pi$. Setting $f(s,a) = A_\pi(s,a)$ immediately gives

$$\mathbb{E}_{\tau\sim\pi'}\Big[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\Big] = \sum_s \rho_{\pi'}(s)\sum_a \pi'(a\mid s)A_\pi(s,a).$$

Substituting this result into (2) yields (1), completing the proof.

3. Why the true objective is hard to optimise

Although (1) is *exact*, it is inconvenient for policy-search algorithms because the state visitation measure $\rho_{\pi'}$ depends on π' in a complicated, recursive way (via the transition dynamics and the entire future sequence of actions). Directly estimating or differentiating it would require on-policy rollouts for *every* candidate policy.

4. A local surrogate that sidesteps the problem

A practical workaround is to *freeze* the visitation frequencies at those of the baseline policy π, producing the *local surrogate objective*

$$L_\pi(\pi') = \eta(\pi) + \sum_s \rho_\pi(s)\sum_a \pi'(a\mid s)A_\pi(s,a). \tag{3}$$

Here only the *action probabilities* $\pi'(a\mid s)$ are optimised, while the weights $\rho_\pi(s)$ and $A_\pi(s,a)$ come from the fixed baseline π. Intuitively, one asks: *"If I visited states exactly as before but could choose different actions, how much would my return change?"*

Next step — trust-region constraint. On its own, (3) is only an approximation; large updates may invalidate the assumption that state frequencies remain unchanged. In the next part of the assignment we will impose a *trust-region* (KL-divergence) constraint that keeps π' close enough to π so that $L_\pi(\pi')$ remains a *reliable* lower bound on $\eta(\pi')$, guaranteeing monotonically improving policies even with finite step sizes.

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 π, π' 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 τ 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 9 and 10 as follows:

$$\eta(\pi') = \eta(\pi) + \mathbb{E}_{\tau\sim\pi'}[\sum_{t=0}^{\infty}\gamma^t \bar{A}(s_t)] \tag{11}$$

$$L_\pi(\pi') = \eta(\pi) + \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty}\gamma^t \bar{A}(s_t)] \tag{12}$$

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

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

c) Prove the following lemma:

Lemma 2.3 *Given that π, π' are α-coupled policies, for all s,*

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

Lemma 2.4 (Coupling bound) *Let (π, π') be an α–*coupled *pair of policies, i.e. for every state s there exists a joint distribution $(a, a') \mid s$ whose marginals are $a \sim \pi(\cdot \mid s)$ and $a' \sim \pi'(\cdot \mid s)$ and $\Pr(a \neq a' \mid s) \leq \alpha$. Define $\bar{A}(s) = \mathop{\mathbb{E}}_{a'\sim\pi'(\cdot|s)} A_\pi(s, a')$. Then for all s*

$$\big|\bar{A}(s)\big| \;\leq\; 2\alpha\,\epsilon, \qquad \textit{where } \epsilon = \max_{s,a}\big|A_\pi(s,a)\big|.$$

Fix a state s and let (a, a') be the coupled action pair at s. Because a has marginal distribution $\pi(\cdot \mid s)$, $\mathbb{E}_a[A_\pi(s,a)] = 0$ by definition of the advantage function (A_π is centred so that its expectation under π vanishes).

1. Express $\bar{A}(s)$ with the coupling. Write the expectation over a' via the joint (a, a'):

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

where the second equality uses $\mathbb{E}_{(a,a')}[A_\pi(s, a)] = \mathbb{E}_a[A_\pi(s, a)] = 0$.

2. Isolate the contribution from disagreement events. If $a = a'$ the integrand is zero, so only the events $\{a \neq a'\}$ matter:

$$\bar{A}(s) = \mathbb{E}_{(a,a')}\big[\mathbf{1}_{\{a\neq a'\}}\,\big(A_\pi(s, a') - A_\pi(s, a)\big)\big].$$

3. Apply absolute values and a uniform bound. Take the absolute value and use $|A_\pi(s, \cdot)| \leq \epsilon$:

$$\big|\bar{A}(s)\big| \leq \mathbb{E}_{(a,a')}\big[\mathbf{1}_{\{a\neq a'\}}\,\big|A_\pi(s, a') - A_\pi(s, a)\big|\big] \leq \mathbb{E}_{(a,a')}\big[\mathbf{1}_{\{a\neq a'\}}\,2\epsilon\big].$$

4. Use the coupling probability. Because the indicator is 1 only when $a \neq a'$,

$$\mathbb{E}_{(a,a')}\big[\mathbf{1}_{\{a\neq a'\}}\big] = \Pr(a \neq a' \mid s) \;\leq\; \alpha,$$

so

$$\big|\bar{A}(s)\big| \leq 2\epsilon\,\alpha.$$

Since s was arbitrary, the inequality holds for every state, proving the lemma.

d) Prove the following lemma:

Lemma 2.5 *Let (π, π') be an α-coupled policy pair. Then:*

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

Lemma 2.6 (One–step drift accumulates geometrically) *Let (π, π') be an α–coupled policy pair, i.e. for every state s there exists a joint law $(a, a')\,|\,s$ such that $\Pr(a \neq a' \mid s) \leq \alpha$ and whose marginals are $a\sim\pi(\cdot \mid s)$, $a'\sim\pi'(\cdot \mid s)$. Then for every horizon $t \geq 0$*

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

Step 1 – Couple whole trajectories. Couple the two rollouts so that

$$s_0 = s'_0, \quad (a_k, a'_k) \mid s_k = s'_k \text{ is the } \alpha\text{-coupling}, \quad s_{k+1}, s'_{k+1} \sim P(\cdot \mid s_k, a_k), P(\cdot \mid s'_k, a'_k)$$

and re-use the *same* transition-noise realisation whenever $a_k = a'_k$. Consequently

$$s_k = s'_k \implies s_{k+1} = s'_{k+1},$$

i.e. the states split only at time indices where the actions disagree.

Step 2 – Probability that states differ at time t. Define the event $E_t = \{s_t \neq s'_t\}$. Because disagreement is only introduced when $a_k \neq a'_k$,

$$\Pr(E_t) \;\leq\; 1 - \prod_{k=0}^{t-1} \Pr(a_k = a'_k \mid E_k = 0) \;\leq\; 1 - (1 - \alpha)^t. \tag{1}$$

Step 3 – Use Lemma 2.4 on $\bar{A}(s)$. From the previous part we have $|\bar{A}(s)| \le 2\alpha\epsilon$ for every state. Hence, for the coupled trajectories,

$$\left|\bar{A}(s'_t) - \bar{A}(s_t)\right| \;\le\; \begin{cases} 0, & \text{if } E_t \text{ does not occur,} \\ |\bar{A}(s'_t)| + |\bar{A}(s_t)| \le 4\alpha\epsilon, & \text{if } E_t \text{ occurs.} \end{cases}$$

Step 4 – Take expectations.

$$\begin{aligned} \left|\mathbb{E}_{s_t\sim\pi'}[\bar{A}(s_t)] - \mathbb{E}_{s_t\sim\pi}[\bar{A}(s_t)]\right| &= \left|\mathbb{E}\big[\bar{A}(s'_t) - \bar{A}(s_t)\big]\right| \\ &\le \mathbb{E}\big[\,|\bar{A}(s'_t) - \bar{A}(s_t)|\,\big] \\ &\le 4\alpha\epsilon\;\Pr(E_t) \\ &\le 4\alpha\epsilon\,\big(1-(1-\alpha)^t\big) \qquad \text{[by (1)],} \end{aligned}$$

which is precisely the stated bound.

e) Prove the following lemma:

Lemma 2.7 *Let (π, π') be an α-coupled policy pair. Then:*

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

Lemma 2.8 (Surrogate–gap bound) *Let (π, π') be an α–coupled pair of policies and define $\epsilon = \max_{s,a}\left|A_\pi(s,a)\right|$. Then*

$$\left|\eta(\pi') - L_\pi(\pi')\right| \;\le\; \frac{4\,\alpha^2\,\gamma\,\epsilon}{(1-\gamma)^2}.$$

Recall the two identities

$$\eta(\pi') \;=\; \eta(\pi) + \sum_{t=0}^{\infty}\gamma^t\,\mathbb{E}_{s_t\sim\pi'}[\bar{A}(s_t)], \qquad L_\pi(\pi') \;=\; \eta(\pi) + \sum_{t=0}^{\infty}\gamma^t\,\mathbb{E}_{s_t\sim\pi}[\bar{A}(s_t)].$$

Hence

$$\left|\eta(\pi') - L_\pi(\pi')\right| \;=\; \left|\sum_{t=0}^{\infty}\gamma^t\left(\mathbb{E}_{s_t\sim\pi'}[\bar{A}(s_t)] - \mathbb{E}_{s_t\sim\pi}[\bar{A}(s_t)]\right)\right| \;\le\; \sum_{t=0}^{\infty}\gamma^t\left|\mathbb{E}_{s_t\sim\pi'}[\bar{A}(s_t)] - \mathbb{E}_{s_t\sim\pi}[\bar{A}(s_t)]\right|.$$

Apply Lemma 2.6. From part (d) we have, for every $t \ge 0$,

$$\left|\mathbb{E}_{s_t\sim\pi'}[\bar{A}(s_t)] - \mathbb{E}_{s_t\sim\pi}[\bar{A}(s_t)]\right| \;\le\; 4\alpha\left(1-(1-\alpha)^t\right)\epsilon.$$

Substituting this bound into the series gives

$$\left|\eta(\pi') - L_\pi(\pi')\right| \;\le\; 4\alpha\epsilon\sum_{t=0}^{\infty}\gamma^t\big(1-(1-\alpha)^t\big).$$

Sum the geometric series exactly. Let $S(\gamma,\alpha) := \sum_{t=0}^{\infty} \gamma^t\big(1-(1-\alpha)^t\big) = \underbrace{\sum_{t=0}^{\infty}\gamma^t}_{=1/(1-\gamma)} - \underbrace{\sum_{t=0}^{\infty}\big[\gamma(1-\alpha)\big]^t}_{=1/\big(1-\gamma(1-\alpha)\big)}$.

A direct subtraction gives

$$S(\gamma,\alpha) = \frac{\gamma\,\alpha}{(1-\gamma)\big((1-\gamma)+\gamma\alpha\big)} \;\le\; \frac{\gamma\,\alpha}{(1-\gamma)^2},$$

because the extra term $\gamma\alpha$ in the second factor of the denominator can only enlarge the denominator.

Combine the pieces. Therefore

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

which is exactly the claimed inequality.

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

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 η:

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

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:

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

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:

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

Theorem 2.9 (Monotonic-improvement bound) *For any two stochastic policies π and π',*

$$\boxed{\eta(\pi') \;\ge\; L_\pi(\pi') \;-\; \frac{4\,\epsilon\,\gamma}{(1-\gamma)^2}\, D_{KL}^{\max}(\pi\,\|\,\pi')} \qquad \big(\;\epsilon \;=\; \max_{s,a}|A_\pi(s,a)|\;\big).$$

The inequality is tight when $\pi' = \pi$.

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

1. Construct an α–coupling from total variation. Let $\alpha := D_{TV}^{\max}(\pi \,\|\, \pi') = \max_s \frac{1}{2}\sum_a \big|\pi(a \mid s) - \pi'(a \mid s)\big|$. By the hint, there exists an α–*coupled* joint law $(a, a') \mid s$ whose marginals are $a \sim \pi(\cdot \mid s)$ and $a' \sim \pi'(\cdot \mid s)$ and that satisfies $\Pr(a \neq a' \mid s) \leq \alpha$ for every state.

2. Apply the surrogate–gap lemma (part (e)). Lemma 2.8 proved that, for any α–coupled pair,

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

Dropping the absolute value gives the desired *lower* bound:

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

3. Convert α^2 to a KL term. For any pair of discrete distributions p, q, Pinsker's inequality gives $D_{TV}(p\,\|\,q)^2 \leq D_{KL}(p\,\|\,q)$. Therefore

$$\alpha^2 = \big(D_{TV}^{\max}(\pi \,\|\, \pi')\big)^2 \;\leq\; D_{KL}^{\max}(\pi \,\|\, \pi').$$

4. Combine. Substituting the KL upper bound for α^2 yields

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

which is exactly the statement of the theorem. The bound becomes an equality when $\pi' = \pi$ because both $D_{KL}^{\max}$ and the surrogate gap vanish.

The theorem guarantees that *any* update $\pi \to \pi'$ satisfying $L_\pi(\pi') - L_\pi(\pi) \;>\; \frac{4\,\epsilon\,\gamma}{(1-\gamma)^2} D_{KL}^{\max}(\pi \,\|\, \pi')$ strictly improves the true return η. Optimising a penalised objective

$$\max_{\pi'} \; L_\pi(\pi') - \frac{4\,\epsilon\,\gamma}{(1-\gamma)^2} D_{KL}^{\max}(\pi \,\|\, \pi')$$

thus yields monotonic performance growth, albeit with very small steps when the theoretical penalty coefficient is used.

In practice one attains larger, yet still reliable, updates by replacing the statewise *maximum* KL with an *average* KL constraint—giving rise to the Trust Region Policy Optimisation (TRPO) algorithm:

$$\begin{aligned} \pi_{k+1} &= \arg\max_{\pi'} \; L_{\pi_k}(\pi') \\ \text{s.t.} \quad & \mathbb{E}_{s \sim \rho_{\pi_k}}\big[D_{KL}\big(\pi_k(\cdot \mid s) \,\|\, \pi'(\cdot \mid s)\big)\big] \;\leq\; \delta. \end{aligned}$$

The average constraint provides a practical surrogate for the stringent statewise bound while preserving the spirit of the theorem: keep each update within a *trust region* where the surrogate objective remains a faithful lower bound on η.