Sharif University of Technology

Masoud Tahmasbi Fard

Student ID: 402200275

CE957: Reinforcement Learning

Assignment #3

May 31, 2024

Table of Contents

1 RL Objective

1.1 Stochasticity of Rewards

In reinforcement learning, the objective is to maximize the expected cumulative reward $J(\theta) = \mathbb{E}\left[\sum_t \gamma^t r(s_t, a_t)\right]$. The use of the expected value in this objective function accounts for the inherent stochasticity in the reinforcement learning process. There are several sources of stochasticity that contribute to the need for using the expected value:

1. **Stochasticity in the Environment:**

   - **State Transitions:** The environment may have probabilistic state transitions, where the next state s_{t+1} is not deterministically determined by the current state s_t and action a_t. Instead, there is a transition probability $P(s_{t+1}|s_t, a_t)$ that defines the likelihood of moving to state s_{t+1} given s_t and a_t.

   - **Reward Function:** The reward function $r(s_t, a_t)$ might also be stochastic, meaning the reward received for a given state-action pair could vary according to some probability distribution $R(r|s_t, a_t)$.

2. **Stochasticity in the Policy:**

   - **Exploration:** In many reinforcement learning algorithms, particularly those using policy gradient methods, the policy $\pi(a_t|s_t, \theta)$ is often stochastic. This stochastic policy is used to ensure sufficient exploration of the state-action space by selecting actions according to some probability distribution rather than deterministically. This helps in learning optimal policies in environments with complex dynamics.

3. **Stochasticity in Initial States:**

   - **Initial State Distribution:** The initial state s_0 from which the agent starts can be drawn from a distribution $P(s_0)$. This introduces variability in the trajectories an agent might experience.

4. **Partial Observability:**

- In some cases, the agent might not have complete information about the environment (e.g., partially observable Markov decision processes - POMDPs). This lack of complete observability adds another layer of stochasticity to the decision-making process, as the agent must rely on belief states or estimates rather than exact states.

5. **Randomness in Sampling:**

- **Experience Replay:** In methods like experience replay used in deep reinforcement learning, mini-batches of experiences are randomly sampled from a replay buffer to update the policy or value function. This introduces additional stochasticity in the learning process.

Due to these sources of stochasticity, the expected value $\mathbb{E}[\cdot]$ is used in the objective function $J(\theta)$. By optimizing the expected cumulative reward, the learning process aims to find a policy that performs well on average, taking into account the various sources of randomness in the environment and the agent's interactions with it. This approach helps in developing robust policies that can handle the variability and uncertainty inherent in real-world scenarios.

1.2 Applying Gradient Descent

If we aim to learn the policy π_θ, there are several challenges in directly using gradient descent to optimize the RL objective:

1. **Why can't we use gradient descent directly on the given equation to optimize the RL objective?**

- **Non-differentiability of the reward function:** The reward function $r(s_t, a_t)$ is typically non-differentiable directly with respect to the policy parameters θ. This makes it challenging to directly compute gradients.

- **Expectation over trajectories:** The objective $J(\theta) = \mathbb{E}\left[\sum_t \gamma^t r(s_t, a_t)\right]$ involves an expectation over all possible trajectories, which is intractable to compute exactly. The

environment dynamics (state transitions) and the policy itself introduce stochasticity, complicating direct gradient computation.

- **Delayed rewards:** In many RL problems, rewards are delayed, meaning the impact of an action might only be observed after several time steps. This temporal aspect makes it difficult to directly associate actions with rewards, complicating gradient calculations.

- **Exploration-exploitation trade-off:** The policy must balance exploration (trying new actions) and exploitation (using known actions that yield high rewards). Direct gradient descent does not inherently address this trade-off.

2. **How can we change the RL setting to be able to optimize the RL objective using gradient descent?** To enable optimization of the RL objective using gradient descent, we can employ the following techniques:

- **Policy Gradient Methods:** These methods reformulate the problem to make it amenable to gradient-based optimization. One commonly used approach is the REINFORCE algorithm, which uses the policy gradient theorem to derive an expression for the gradient of the expected reward:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_t \nabla_\theta \log \pi_\theta(a_t|s_t) \sum_{t'} \gamma^{t'-t} r(s_{t'}, a_{t'})\right]$$

  In this new objective, we calculate the expected value with respect to the policy π_θ instead of all sources of randomness in the original objective. By selecting actions based on a parametrized policy, gaining rewards, and creating trajectories, we can apply gradient descent optimization. This expression can be estimated using samples from the policy, allowing the use of stochastic gradient ascent.

- **Actor-Critic Methods:** These methods combine policy-based and value-based approaches. The actor updates the policy parameters θ via gradient ascent, while the critic estimates the value function $V^\pi(s)$ or the action-value function $Q^\pi(s, a)$. The gradient

used for updating the actor can be written as:

$$\nabla_\theta J(\theta) \approx \mathbb{E}_{\pi_\theta} \left[\nabla_\theta \log \pi_\theta(a_t|s_t)(Q^\pi(s_t, a_t) - b(s_t)) \right]$$

where $b(s_t)$ is a baseline that reduces variance. Similar to policy gradient methods, we calculate the expected value with respect to the policy, allowing us to apply gradient descent optimization by selecting actions, gaining rewards, and creating trajectories.

- **Proximal Policy Optimization (PPO):** PPO is a more advanced method that uses a clipped surrogate objective to maintain a balance between exploration and exploitation while ensuring stable policy updates. The PPO objective can be optimized using gradient descent:

$$L^{CLIP}(\theta) = \mathbb{E}_t \left[\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t) \right]$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$. Here, the expected value is calculated with respect to the policy, enabling gradient descent optimization by following the same principle of selecting actions, gaining rewards, and creating trajectories.

By employing these techniques, we transform the RL problem into a form that allows the application of gradient descent and other optimization methods, facilitating the learning of optimal policies in complex, stochastic environments.

2 Bias & Variance of Policy Gradient

2.1 Unbiased Estimation of the RL Objective

To prove that the given estimation of $\nabla_\theta J(\theta)$ is an unbiased estimation, we need to show that its expected value is equal to the true gradient $\nabla_\theta J(\theta)$.

The estimation given is:

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \left(\sum_t \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}) \right) \left(\sum_t \gamma^t r(s_{i,t}, a_{i,t}) \right)$$

We will start by expressing the true policy gradient using the policy gradient theorem:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_t \nabla_\theta \log \pi_\theta(a_t|s_t) \sum_{t'} \gamma^{t'} r(s_t, a_{t'}) \right]$$

where τ denotes a trajectory $(s_0, a_0, s_1, a_1, \ldots)$ sampled from the policy π_θ.

Now, consider the given estimator:

$$\hat{\nabla}_\theta J(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left(\sum_t \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}) \right) \left(\sum_t \gamma^t r(s_{i,t}, a_{i,t}) \right)$$

To show that this is an unbiased estimator, we need to demonstrate that:

$$\mathbb{E}[\hat{\nabla}_\theta J(\theta)] = \nabla_\theta J(\theta)$$

We assume that each trajectory $\tau_i = (s_{i,0}, a_{i,0}, s_{i,1}, a_{i,1}, \ldots)$ is sampled independently according to the policy π_θ. Therefore, the expectation of the estimator can be written as:

$$\mathbb{E}[\hat{\nabla}_\theta J(\theta)] = \mathbb{E} \left[\frac{1}{N} \sum_{i=1}^{N} \left(\sum_t \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}) \right) \left(\sum_t \gamma^t r(s_{i,t}, a_{i,t}) \right) \right]$$

By the linearity of expectation, we can move the expectation inside the sum:

$$\mathbb{E}[\hat{\nabla}_\theta J(\theta)] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left[\left(\sum_t \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}) \right) \left(\sum_t \gamma^t r(s_{i,t}, a_{i,t}) \right) \right]$$

Since the trajectories are sampled independently and identically, the expectation for each trajectory is the same:

$$\mathbb{E}[\hat{\nabla}_\theta J(\theta)] = \mathbb{E} \left[\left(\sum_t \nabla_\theta \log \pi_\theta(a_t|s_t) \right) \left(\sum_t \gamma^t r(s_t, a_t) \right) \right]$$

Next, we use the fact that the expectation of a product of random variables, where one is the gradient of the log-policy, can be rewritten using the policy gradient theorem:

$$\mathbb{E}\left[\left(\sum_t \nabla_\theta \log \pi_\theta(a_t|s_t)\right)\left(\sum_t \gamma^t r(s_t,a_t)\right)\right] = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_t \nabla_\theta \log \pi_\theta(a_t|s_t)\sum_{t'}\gamma^{t'} r(s_{t'},a_{t'})\right]$$

Therefore, we have:

$$\mathbb{E}[\hat{\nabla}_\theta J(\theta)] = \nabla_\theta J(\theta)$$

The key assumption in the proof is that the trajectories $\tau_i = (s_{i,0}, a_{i,0}, s_{i,1}, a_{i,1}, \ldots)$ are independently and identically distributed (i.i.d.) according to the policy π_θ. This means each trajectory is generated independently from the same policy π_θ, and each state-action-reward sequence is sampled in the same way.

In another view, the given estimator is a Monte Carlo estimation of the desired expected value. As long as the samples are i.i.d., Monte Carlo estimation is known to be an unbiased estimation. Hence, the assumption that trajectories are i.i.d. ensures that our Monte Carlo estimator for $\nabla_\theta J(\theta)$ is unbiased.

2.2 Q-Function in PG Method

The Policy Gradient Theorem (Causality Trick) states that:

$$\begin{aligned}\nabla_\theta J(\theta) &= \mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\nabla_\theta \log \pi_\theta(a_t|s_t)\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right]\\ &= \mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\nabla_\theta \log \pi_\theta(a_t|s_t)\sum_{t'=t}^{\infty}\gamma^{t'-t} r_{t'}\right]\end{aligned}$$

where $r_{t'} = r(s_{t'}, a_{t'})$.

if we define $\rho_\pi(s) = \sum_{t=0}^{\infty}\gamma^t p(s_t = s)$, and $\hat{Q}(s_t,a_t) = \sum_{t'=t}^{\infty}\gamma^{t'-t} r_{t'}$, we can rewrite the above equation (with abuse of notation) as:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\hat{Q}(s_t,a_t)\right].$$

2.3 Baseline

We want to derive the optimal state-dependent baseline, which minimizes the variance of the policy gradient estimate. More precisely, we minimize the trace of the covariance of the policy gradient; that is, the sum of the variance of the components of the vectors.

$$\nabla_\theta J(\theta) := \mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\left(\hat{Q}(s_t,a_t) - b(s_t)\right)\right]$$

Let g be the associated random variable, that is, $\nabla_\theta J(\theta) = \mathbb{E}_{\rho_\pi,\pi}[g]$:

$$g := \nabla_\theta \log \pi_\theta(a_t|s_t)\left(\hat{Q}(s_t,a_t) - b(s_t)\right), \quad a_t \sim \pi_\theta(a_t|s_t), s_t \sim \rho_\pi(s_t)$$

The variance of the policy gradient is:

$$\begin{aligned}
\mathrm{Var}(g) &= \mathbb{E}_{\rho_\pi,\pi}\left[(g - \mathbb{E}_{\rho_\pi,\pi}[g])^T(g - \mathbb{E}_{\rho_\pi,\pi}[g])\right] \\
&= \mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\right] b(s_t)^2 \\
&\quad - 2\mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\hat{Q}(s_t,a_t)\right] b(s_t)
\end{aligned}$$

Note that $\mathbb{E}[J(\theta)]$ contains a bias-free term, by the score function argument, which then does not affect the minimizer. Terms which do not depend on $b(s_t)$ also do not affect the minimizer.

$$\begin{aligned}
\frac{\partial}{\partial b}[\mathrm{Var}(g)] &= 0 \\
&= 2\mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\right] b(s_t) \\
&\quad - 2\mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\hat{Q}(s_t,a_t)\right] \\
\implies b^*(s_t) &= \frac{\mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\hat{Q}(s_t,a_t)\right]}{\mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\right]}
\end{aligned}$$

2.4 Optimal Q-Function

We will assume that given π_θ, $\nabla_\theta \log \pi_\theta(a_t|s_t)^T$, $\nabla_\theta \log \pi_\theta(a_t|s_t)$, and $Q^{\pi_\theta}(s_t,a_t)$ are independent random variables:

$$\begin{aligned}
b(s_t) &= \frac{\mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T\right]\mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\right]\mathbb{E}_{\rho_\pi,\pi}\left[Q^{\pi_\theta}(s_t,a_t)\right]}{\mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T\right]\mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\right]} \\
&= \mathbb{E}_{\rho_\pi,\pi}\left[Q^{\pi_\theta}(s_t,a_t)\right] = V^\pi(s_t)
\end{aligned}$$

3 Trust Region Policy Optimization

3.1 Bounding the Distribution Change (Part a & b)

We want to prove that the state distribution $p_\theta(s_t)$ under policy π_θ is close to the state distribution $p_{\theta'}(s_t)$ under a nearby policy $\pi_{\theta'}$. Let's consider two cases:

Case 1: Deterministic Policy Assume π_θ is a deterministic policy, *i.e.*, $a_t = \pi_\theta(s_t)$. We say $\pi_{\theta'}$ is close to π_θ if the probability of taking a different action under $\pi_{\theta'}$ is small: $\pi_{\theta'}(a_t \neq \pi_\theta(s_t)|s_t) \leq \epsilon$. Then, we can show:

$$
\begin{aligned}
p_{\theta'}(s_t) &= (1-\epsilon)^t p_\theta(s_t) + (1-(1-\epsilon)^t) p_{\text{mistake}}(s_t) \\
|p_{\theta'}(s_t) - p_\theta(s_t)| &= (1-(1-\epsilon)^t)|p_{\text{mistake}}(s_t) - p_\theta(s_t)| \\
&\leq 2(1-(1-\epsilon)^t) \\
&\leq 2t\epsilon \text{ since } (1-\epsilon)^t \geq 1 - t\epsilon \text{ for } \epsilon \in [0,1]
\end{aligned}
$$

Case 2: Stochastic Policy Now, consider the case where π_θ is a stochastic policy. We say $\pi_{\theta'}$ is close to π_θ if the difference in action probabilities is small: $|\pi_{\theta'}(a_t|s_t) - \pi_\theta(a_t|s_t)| \leq \epsilon$ for all s_t. Using a useful lemma, we can show that $\pi_{\theta'}(a_t|s_t)$ takes a different action than $\pi_\theta(a_t|s_t)$ with probability at most ϵ.

Lemma 1.

$$
\begin{aligned}
&\text{if } |p_X(x) - p_Y(x)| = \epsilon, \exists p(x,y) \text{ s.t. } p(x) = p_X(x), p(y) = p_Y(y), p(x=y) = 1-\epsilon \\
&\qquad \implies p_X(x) \text{ "agrees" with } p_Y(y) \text{ w.p. } 1-\epsilon \\
&\qquad \implies \pi_{\theta'}(a_t|s_t) \text{ takes a different action than } \pi_\theta(a_t|s_t) \text{ w.p. at most } \epsilon
\end{aligned}
$$

Thus we can show that

$$
\begin{aligned}
|p_{\theta'}(s_t) - p_\theta(s_t)| &= (1-(1-\epsilon)^t)|p_{\text{mistake}}(s_t) - p_\theta(s_t)| \\
&\leq 2(1-(1-\epsilon)^t) \\
&\leq 2t\epsilon
\end{aligned}
$$

In both cases, we have shown that if $\pi_{\theta'}$ is close to π_θ, then the state distributions $p_{\theta'}(s_t)$ and $p_\theta(s_t)$ are also close, with a difference bounded by $2t\epsilon$. Now, consider a function $f(\cdot)$ and its expectation under the two distributions:

$$\mathbb{E}_{s_t\sim p_{\theta'}(s_t)}[f(s_t)] \geq \mathbb{E}_{s_t\sim p_\theta(s_t)}[f(s_t)] - 2t\epsilon \max_{s_t} f(s_t)$$

Proof:

$$\begin{aligned}
\mathbb{E}_{s_t\sim p_{\theta'}(s_t)}[f(s_t)] &= \sum_{s_t} p_{\theta'}(s_t) f(s_t) \\
&\geq \sum_{s_t} p_\theta(s_t) f(s_t) - |p_\theta(s_t) - p_{\theta'}(s_t)| \max_{s_t} f(s_t) \\
&\geq \mathbb{E}_{s_t\sim p_\theta(s_t)}[f(s_t)] - 2t\epsilon \max_{s_t} f(s_t)
\end{aligned}$$

Using this result, we can show that maximizing a certain expectation with respect to θ' will also maximize (up to a bounded error) the same expectation with respect to θ, which is related to the RL objective:

$$\begin{aligned}
&\sum_t \mathbb{E}_{s_t\sim p_{\theta'}(s_t)} \left[\mathbb{E}_{a_t\sim\pi_\theta(a_t|s_t)} \left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t A_{\pi_\theta}(s_t, a_t) \right] \right] \\
&\geq \sum_t \mathbb{E}_{s_t\sim p_\theta(s_t)} \left[\mathbb{E}_{a_t\sim\pi_\theta(a_t|s_t)} \left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t A_{\pi_\theta}(s_t, a_t) \right] \right] - \sum_t 2t\epsilon C
\end{aligned}$$

where $C = O(Tr_{\max})$ if finite and $C = O(\frac{r_{\max}}{1-\gamma})$ if infinite.

Proof:

Finite Case: Let's assume the maximum reward $r_{\max}$ is finite. We want to show that $C = O(Tr_{\max})$.

$$\begin{aligned}
\max_{s_t} f(s_t) &= \max_{s_t} \mathbb{E}_{a_t\sim\pi_\theta(a_t|s_t)} \left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t A_{\pi_\theta}(s_t, a_t) \right] \\
&\leq \max_{s_t} \mathbb{E}_{a_t\sim\pi_\theta(a_t|s_t)} \left[\gamma^t r_{\max} \right] \\
&= r_{\max} \sum_{t=0}^{T} \gamma^t \\
&\leq r_{\max} \sum_{t=0}^{T} 1 \\
&= T r_{\max}
\end{aligned}$$

Since $\sum_t 2t\epsilon \max_{s_t} f(s_t) \leq \sum_t 2t\epsilon r_{\max}$, we have $C = O(Tr_{\max})$.

Infinite Case: Now, let's assume the maximum reward $r_{\max}$ is infinite. We want to show that $C = O(\frac{r_{\max}}{1-\gamma})$.

$$\begin{aligned}
\max_{s_t} f(s_t) &= \max_{s_t} \mathbb{E}_{a_t \sim \pi_\theta(a_t|s_t)} \left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t A_{\pi_\theta}(s_t, a_t) \right] \\
&\leq \max_{s_t} \mathbb{E}_{a_t \sim \pi_\theta(a_t|s_t)} \left[\gamma^t r_{\max} \right] \\
&= r_{\max} \sum_{t=0}^{\infty} \gamma^t \\
&= \frac{r_{\max}}{1-\gamma}
\end{aligned}$$

Since $\sum_t 2t\epsilon \max_{s_t} f(s_t) \leq \sum_t 2t\epsilon \frac{r_{\max}}{1-\gamma}$, we have $C = O(\frac{r_{\max}}{1-\gamma})$.

Therefore, maximizing the left-hand side expectation with respect to θ', subject to the constraint $|\pi_{\theta'}(a_t|s_t) - \pi_\theta(a_t|s_t)| \leq \epsilon$ for small enough ϵ, is guaranteed to improve the RL objective $J(\theta') - J(\theta)$. Thus, the summarized update rule would be:

$$\theta' \leftarrow \arg\max_{\theta'} \sum_t \mathbb{E}_{s_t \sim p_\theta(s_t)} \left[\mathbb{E}_{a_t \sim \pi_\theta(a_t|s_t)} \left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t A_{\pi_\theta}(s_t, a_t) \right] \right]$$

3.2 Policy Gradients with Constraints (Part c)

Assuming that $|\pi_{\theta'}(a_t|s_t) - \pi_\theta(a_t|s_t)| \geq \epsilon$, we aim to find a lower bound for $D_{KL}(\pi_{\theta'}(a_t|s_t) \| \pi_\theta(a_t|s_t))$. Using the Pinsker's inequality: $|\pi_{\theta'}(a_t|s_t) - \pi_\theta(a_t|s_t)| \leq \sqrt{\frac{1}{2} D_{KL}(\pi_{\theta'}(a_t|s_t) \| \pi_\theta(a_t|s_t))}$ Therefore:

$$\sqrt{\frac{1}{2} D_{KL}(\pi_{\theta'}(a_t|s_t) \| \pi_\theta(a_t|s_t))} \geq \epsilon \implies D_{KL}(\pi_{\theta'}(a_t|s_t) \| \pi_\theta(a_t|s_t)) \geq 2\epsilon^2$$

3.3 Taylor Expansion (Part d)

Objective:

$$\theta' \leftarrow \arg\max_{\theta'} \sum_t \mathbb{E}_{s_t \sim p_\theta(s_t)} \left[\mathbb{E}_{a_t \sim \pi_\theta(a_t|s_t)} \left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t A^{\pi_\theta}(s_t, a_t) \right] \right] = \arg\max_{\theta'} \bar{A}(\theta')$$

Now we want to use first order Taylor approximation for the objective:

$$\theta' \leftarrow \arg\max_{\theta'} \nabla_\theta \bar{A}(\theta)^\top (\theta' - \theta)$$

such that $D_{\mathrm{KL}}(\pi_{\theta'}(a_t|s_t)\|\pi_\theta(a_t|s_t)) \leq \epsilon$. Applying the log-trick:

$$\begin{aligned}\nabla_\theta \bar{A}(\theta) &= \sum_t \mathbb{E}_{s_t\sim p_\theta(s_t)}\left[\mathbb{E}_{a_t\sim\pi_\theta(a_t|s_t)}\left[\gamma^t \nabla_\theta \log \pi_\theta(a_t|s_t) A^{\pi_\theta}(s_t,a_t)\right]\right] \\ &= \nabla_\theta J(\theta)\end{aligned}$$

which is exactly the normal policy gradient. So our objective is:

$$\theta' \leftarrow \arg\max_{\theta'} \nabla_\theta J(\theta)^\top (\theta' - \theta)$$

such that $D_{\mathrm{KL}}(\pi_{\theta'}(a_t|s_t)\|\pi_\theta(a_t|s_t)) \leq \epsilon$.

4 Deterministic Policy Gradient

4.1 Proof

The proof follows a similar approach to the standard stochastic policy gradient theorem. The regularity conditions ensure that the value function $V^{\mu_\theta}(s)$ and its gradient $\nabla_\theta V^{\mu_\theta}(s)$ are continuous with respect to θ and s. Additionally, the compactness of the state space S implies that for any policy parameter θ, the norms $\|\nabla_\theta V^{\mu_\theta}(s)\|$, $\|\nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}\|$, and $\|\nabla_\theta \mu_\theta(s)\|$ are bounded functions of s. These conditions enable exchanging derivatives and integrals, as well as the order of integration when necessary in the proof.

We begin by expressing the gradient of the value function:

$$
\begin{aligned}
\nabla_\theta V^{\mu_\theta}(s) &= \nabla_\theta Q^{\mu_\theta}(s, \mu_\theta(s)) \\
&= \nabla_\theta \left(r(s, \mu_\theta(s)) + \int_S \gamma p(s'|s, \mu_\theta(s)) V^{\mu_\theta}(s') ds' \right) \\
&= \nabla_\theta \mu_\theta(s) \nabla_a r(s,a)|_{a=\mu_\theta(s)} + \nabla_\theta \int_S \gamma p(s'|s, \mu_\theta(s)) V^{\mu_\theta}(s') ds' \\
&= \nabla_\theta \mu_\theta(s) \nabla_a r(s,a)|_{a=\mu_\theta(s)} + \\
&\quad \int_S \gamma \left(p(s'|s, \mu_\theta(s)) \nabla_\theta V^{\mu_\theta}(s') + \nabla_\theta \mu_\theta(s) \nabla_a p(s'|s,a)|_{a=\mu_\theta(s)} V^{\mu_\theta}(s') \right) ds' \qquad (1) \\
&= \nabla_\theta \mu_\theta(s) \nabla_a \left(r(s,a) + \int_S \gamma p(s'|s,a) V^{\mu_\theta}(s') ds' \right) \Big|_{a=\mu_\theta(s)} + \int_S \gamma p(s'|s, \mu_\theta(s)) \nabla_\theta V^{\mu_\theta}(s') ds' \\
&= \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} + \int_S \gamma p(s \to s', 1, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s') ds'.
\end{aligned}
$$

In the step marked (1), we applied the Leibniz integral rule to exchange the order of differentiation and integration. This step requires the regularity conditions, specifically the continuity of $p(s'|s,a)$, $\mu_\theta(s)$, $V^{\mu_\theta}(s)$, and their derivatives with respect to θ. By iterating this formula, we

obtain:

$$
\begin{aligned}
&= \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} + \int_S \gamma p(s \to s', 1, \mu_\theta) \nabla_\theta \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')} ds' \\
&\quad + \int_S \gamma p(s \to s', 1, \mu_\theta) \int_S \gamma p(s' \to s'', 1, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s'') ds'' ds' \\
&= \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} + \int_S \gamma p(s \to s', 1, \mu_\theta) \nabla_\theta \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')} ds' \\
&\quad + \int_S \gamma^2 p(s \to s', 2, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s') ds' \qquad (2) \\
&\dots \\
&= \int_S \sum_{t=0}^{\infty} \gamma^t p(s \to s', t, \mu_\theta) \nabla_\theta \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')} ds'.
\end{aligned}
$$

In the step marked (2), we used Fubini's theorem to exchange the order of integration, requiring the regularity conditions that ensure $\|\nabla_\theta V^{\mu_\theta}(s)\|$ is bounded. Next, we take the expectation over the initial state distribution S_1:

$$
\begin{aligned}
\nabla_\theta J(\mu_\theta) &= \nabla_\theta \int_S p_1(s) V^{\mu_\theta}(s) ds \\
&= \int_S p_1(s) \nabla_\theta V^{\mu_\theta}(s) ds \qquad (3) \\
&= \int_S \int_S \sum_{t=0}^{\infty} \gamma^t p_1(s) p(s \to s', t, \mu_\theta) \nabla_\theta \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')} ds' ds \\
&= \int_S \rho_\mu(s) \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} ds,
\end{aligned}
$$

In step (3), we applied the Leibniz integral rule to exchange the derivative and integral, requiring the regularity conditions that ensure $p_1(s)$, $V^{\mu_\theta}(s)$, and their derivatives with respect to θ are continuous. In the last line, we again used Fubini's theorem to exchange the order of integration, which is permitted due to the boundedness of the integrand implied by the regularity conditions. Thus, we have prooved that:

$$
\nabla_\theta J(\mu_\theta) = \int_S \rho_\mu(s) \nabla_\theta \mu_\theta(S) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} ds = \mathbb{E}_{s \sim \rho_\mu} \left[\nabla_\theta \mu_\theta(S) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} \right]
$$