REINFORCEMENT LEARNING

JUBAYER IBN HAMID

Contents

Preface 2
1. Reinforcement Learning Framework 3
1.1. Value Functions and Optimal Policies 5
1.2. Bellman Operators 8
1.3. Optimal Policy 11
1.4. State Distributions 12
2. Policy Gradient Methods 14
2.1. Policy Gradient 15
2.2. REINFORCE 17
2.3. On-Policy Actor-Critic Methods 23
2.4. Performance Difference Lemma 25
References 27
Appendix A. Banach Fixed-Point Theorem 28

Preface

This document is a combination of notes I wrote for CS 224R (taught by Prof. Chelsea Finn in Spring, 2025) as the head teaching assistant. The notes contain material taught in CS 234 taught by Prof. Emma Brunskill at Stanford University, notes from various textbooks and papers (in particular, Richard Sutton and Andrew Barto's "Reinforcement Learning: An Introduction" [SB18]).

Although these are reading notes, there may be various errors throughout, both minor and major, which almost certainly did not appear in the original works that I was reading. If you find any, please let me know by sending an email at jubayer@stanford.edu. I have also not been able to stick to the same notation throughout these notes and I apologize for that - in my defence, reinforcement learning is notorious for a lack of universal notation. Hopefully, the inconsistencies are not too troublesome.

1. Reinforcement Learning Framework

Reinforcement learning is a framework in which an agent acts in an environment where it gets to sequentially choose actions over time steps and observes feedback from the environment along the way. The goal of the agent is to maximize cumulative reward. This is typically studied as a Markov decision process (MDP), which is a mathematical formalization of sequential decision-making; in this framework, actions influence both the rewards received and the states subsequently observed. The learner is referred to as the agent (or, sometimes, the model). The agent interacts with an environment at each discrete time step $t = 0, 1, 2, 3, \cdots$. At each time step t, the agent observes a state s_t, which represents the environment at that time. We denote the state space by $\mathcal{S}$, which is the set of all possible states, and denote the state S_t as a random variable taking values in $\mathcal{S}$. The agent conditions on its observation $S_t = s_t$ and chooses an action a_t. We denote the action space by $\mathcal{A}$, which is the set of all possible actions, and denote A_t as a random variable taking values in $\mathcal{A}$. As a consequence of taking action a_t, the agent receives a reward r_t, which takes values in the reward space $\mathcal{R}$, and the environment transitions to the next state s_{t+1}. Thus, a trajectory takes the form

$$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots).$$

A defining characteristic of an MDP is the Markov property, which states that the conditional distribution of the next state and reward depends only on the current state and action, rather than the entire preceding history. This means the state s_t anad the action a_t constitute sufficient information for the environment's distribution over the next state and reward to be determined. Formally,

$$P(s_{t+1}, r_t \mid s_t, a_t, s_{t-1}, a_{t-1}, \ldots) = P(s_{t+1}, r_t \mid s_t, a_t).$$

In particular, marginalizing over rewards gives

$$P(s_{t+1} \mid s_t, a_t, s_{t-1}, a_{t-1}, \ldots) = P(s_{t+1} \mid s_t, a_t).$$

The conditional distribution $P(\cdot \mid s_t, a_t)$ is referred to as the transition dynamics of the environment.

A deterministic environment is one in which the transition dynamics is deterministic, i.e. $P(S_{t+1} \mid s_t, a_t) = 1\{S_{t+1} = s\}$ for some s. Otherwise, the environment is called stochastic. We typically assume that the transition dynamics are time-homogeneous or stationary, meaning that $P(s_{t+1} \mid s_t, a_t)$ does not depend explicitly on the time step t; i.e. $P(S_{t+1} = s' \mid S_t = s, A_t = a) = P(S_{t'+1} = s' \mid S_{t'} = s, A_{t'} = a)$ for any t and t'. Finally, a finite MDP is an MDP in which the state space $\mathcal{S}$ and action space $\mathcal{A}$ are both finite. Throughout these notes, we will work with finite MDPs, as they are the most practically relevant cases.

Technically, one can have a distribution over rewards at any given time step, i.e. $\mathbb{P}(R_t = r_t \mid S_t = s_t, A_t = a_t)$. We will usually refer to $r(s_t, a_t)$ as the reward at time t. If the reward function is deterministic, then this is simply the reward attained at time t after taking action a_t from state s_t. If the reward function is stochastic, we will write $r(s_t, a_t)$ to mean the expected reward at time t, i.e. $r(s_t, a_t) := \mathbb{E}[R(s_t, a_t) \mid S_t = s_t, A_t = a_t] = \sum_{r \in \mathcal{R}} \mathbb{P}(R_t = r \mid S_t = s_t, A_t = a_t) r$. (we will always assume that the conditional reward distribution is stationary as well). In either case, we can simply assume that given $S_t = s_t$ and $A_t = a_t$, our agent receives some reward $r(s_t, a_t)$ as the reward and not worry about whether or not the reward function is deterministic.

We say the agent executes actions by following a *policy* which is a distribution π_θ. The agent takes actions in the sense that we sample

$$a_t \sim \pi_\theta(\cdot \mid S_t = s_t)$$

at each time step t. Note that, over a trajectory, the agent observes rewards as $R_0, R_1, R_2, \cdots$. In reinforcement learning, the goal of our agent is to find the optimal policy, i.e. the policy that maximizes the expected sum of rewards.

We formalize this goal more precisely now. We will refer to the sum of rewards from a given time step t onwards as the return, denoted G_t, i.e. $G_t = R_t + R_{t+1} + R_{t+2} + \cdots$. In some environments, there is a notion of an *episode* where the task that the agent is attempting to execute ends at some time step (such as a game of football). In that case, the return at time t is $G_t = R_t + R_{t+1} + R_{t+2} + \cdots + R_T$. We say that an episode ends at a terminal state after

which the environment is reset to an initial state. We refer to such tasks as episodic tasks and the time horizon as the maximum time step that an episode can go up to. In other tasks, one may not be able to establish a notion of episodes where the task ultimately never ends. Such tasks are referred to as continuing tasks or infinite horizon tasks. In almost all applications, we will be looking at episodic tasks although we will study algorithms from both points of view.

One can immediately observe that in continuing tasks, our definition of the return at time step t has a problem. If we set $G_t := R_t + R_{t+1} + R_{t+2} + \cdots$, then G_t could diverge to infinity (for e.g., suppose each $R_{t'} = 1$, then the sum of all the rewards blows up to infinity). As such, in continuing tasks, we discount the reward at each time step. Let $\gamma \in [0, 1)$ be the discount factor. We define the return to be

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k}.$$

This allows us to bypass the problem of diverging returns since, if the reward at any given time step is bounded within $[R_l, R_u]$, then define $R_{\max} = \max\{|R_l|, |R_u|\}$ and we have that

$$\begin{aligned} |G_t| &= \left| \sum_{k=0}^{\infty} \gamma^k R_{t+k} \right| \\ &\leq \sum_{k=0}^{\infty} \left| \gamma^k R_{t+k} \right| \\ &\leq \sum_{k=0}^{\infty} \gamma^k \left| R_{t+k} \right| \\ &\leq \sum_{k=0}^{\infty} \gamma^k R_{\max} \\ &\leq R_{\max} \cdot \frac{1}{1-\gamma}. \end{aligned}$$

The interpretation of the discount factor is that we are imposing the requirement that short term rewards are more important than longer term ones. The value of γ determines how strongly we prefer short term rewards over longer term ones. Observe that we can write the return as

$$G_t = R_t + \gamma G_{t+1}.$$

Also note that discounting can be used even in episodic tasks and our definition of return under discounting works in that setting too by simply setting $R_{t'}$ to be 0 for all $t' > T$ where T is the time horizon.

With this notational convention, we can succintly write that the goal for an agent in reinforcement learning is to learn the policy π_θ that maximizes the expected discounted sum of reward:

$$\max_{\pi_\theta} \mathbf{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]. \tag{1}$$

Here, we write $\mathbf{E}_\pi$ to emphasize that in the trajectories we are sampling, each A_t is sampled as $a_t \sim \pi_\theta(\cdot \mid S_t = s_t)$, while the states are sampled as $s_{t+1} \sim P(\cdot \mid S_t = s_t, A_t = a_t)$. We sample the initial state S_0 from some initial state distribution which we will simply denote by $P(S_0 = s_0)$.

1.1. **Value Functions and Optimal Policies**

A central construction in reinforcement learning is the value function. Quite often we are interested in understanding the value of a state in the sense that if a policy π starts from state s, what is the expected sum of rewards the policy will attain? In this case, the value function computes the expected sum of rewards starting from state s and then sampling actions under policy π thereafter.

Definition 1.1 (State-value function)**.** *The state-value function of a state s under policy π is defined to be the expected return starting from state s and following policy π thereafter. In continuing tasks, this is written as*

$$V_\pi(s) = \mathbf{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\Big|S_0=s\right] = \mathbf{E}_\pi\left[\sum_{t=t'}^{\infty}\gamma^{t-t'} r(s_t,a_t)\Big|S_{t'}=s\right], \quad \forall \text{ time step } t'. \tag{2}$$

Note that, in continuing tasks, the value function does not depend on the time step at which state s has been visited since the transition dynamics distribution is stationary and we are assuming the policy is stationary as well; one simply needs to consider the trajectory from whichever time step our agent visits state s at onwards.

In episodic tasks of horizon T, we write this as

$$V_\pi(s) = \mathbf{E}_\pi\left[\sum_{t=0}^{T}\gamma^t r(s_t,a_t)\Big|S_0=s\right].$$

The state-value function of state s at time t is the expected return starting from state s being visited at time t specifically and following policy π thereafter:

$$V_\pi(s;t) = \mathbf{E}_\pi\left[\sum_{k=t}^{T}\gamma^{k-t} r(s_k,a_k) \mid S_t=s\right]$$

⌟

Here, the expectation is over trajectories sampled under π, i.e. we sample

$$\tau = [s_0,a_0,r_0,s_1,a_1,r_1,\cdots] \sim P^\pi.$$

Note that the probability of sampling this specific trajectory is:

$$P_\theta(\tau) = \rho(s_0)\prod_{i=0}^{\infty}\pi(a_i|s_i)P(s_{i+1}|s_i,a_i)$$

in the continuing task setting and in the finite horizon setting, if the trajectory has T time steps, then it is:

$$P_\theta(\tau) = \rho(s_0)\prod_{i=0}^{T-1}\pi(a_i|s_i)P(s_{i+1}|s_i,a_i).$$

Remark 1.2. Note that the state-value function conditions on the fact that the policy is starting from state s. This state could very well be one that policy π would not visit unconditionally. The purpose of asking for the state-value function of s is to essentially ask what happens *if* the policy did, somehow, start from state s and started acting from that state onwards. ⌟

Remark 1.3. Here's a short proof of why (2) holds. Let $t' \geq 0$ be arbitrary and define $k = t - t'$. Then, the right-hand side in (2) becomes

$$\mathbf{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k r(S_{t'+k},A_{t'+k})\Big|S_{t'}=s\right].$$

Since the transition dynamics are time-homogeneous and the policy π is stationary and Markov, we have

$$\mathbb{P}_\pi(S_{t'+k}=s', A_{t'+k}=a \mid S_{t'}=s) = \mathbb{P}_\pi(S_k=s', A_k=a \mid S_0=s)$$

for every $k \geq 0$. This follows from the Markov property: conditional on the present state, the distribution of future states and actions depends only on that state, not on the time at which it was visited or the history preceding it. Therefore,

$$\begin{aligned}\mathbf{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k r(S_{t'+k}, A_{t'+k})\Big| S_{t'} = s\right] &= \sum_{k=0}^{\infty}\gamma^k \mathbf{E}_\pi\left[r(S_{t'+k}, A_{t'+k}) \mid S_{t'} = s\right] \\ &= \sum_{k=0}^{\infty}\gamma^k \mathbf{E}_\pi\left[r(S_k, A_k) \mid S_0 = s\right] \\ &= \mathbf{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k r(S_k, A_k) \mid S_0 = s\right] \\ &= V_\pi(s).\end{aligned}$$

Here, we assume that rewards are bounded and $\gamma \in [0, 1)$, which justifies exchanging the infinite summation and conditional expectation. We also assume that $\mathbb{P}_\pi(S_{t'} = s) > 0$, so that the conditional expectations are well-defined. ⌟

Definition 1.4 (Action-value function)**.** *The action-value function of a state-action pair (s, a) under policy π is defined to be the expected return starting from state s, taking action a, and following policy π thereafter. In continuing tasks, this is written as*

$$Q_\pi(s, a) = \mathbf{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)\Big| S_0 = s, A_0 = a\right] = \mathbf{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)\Big| S_{t'} = s, A_{t'} = a\right], \quad \forall t'. \tag{3}$$

Again, the time step at which the policy takes action a from state s does not matter since the transition dynamics distribution and the policy are stationary (see remark above).

In episodic tasks of horizon T, we write this as

$$Q_\pi(s, a) = \mathbf{E}_\pi\left[\sum_{t=0}^{T}\gamma^t r(s_t, a_t)\Big| S_0 = s, A_0 = a\right].$$

The action-value function of state-action pair (s, a) at time t is the expected return starting from state s being visited at time t, taking action a at that time, and following policy π thereafter:

$$Q_\pi(s, a; t) = \mathbf{E}_\pi\left[\sum_{k=t}^{T}\gamma^{k-t} r(s_k, a_k)\Big| S_t = s, A_t = a\right].$$

⌟

Proposition 1.5 (Bellman Equation for state-value function)**.** *For any stationary policy π and any state s, the Bellman equation for continuing tasks is*

$$V_\pi(s) = \mathbf{E}_{a\sim\pi(\cdot|s), s'\sim P(\cdot|s,a)}\left[r(s, a) + \gamma V_\pi(s')\right].$$

For finite-horizon episodic tasks with horizon T, the Bellman equation is

$$V_\pi(s; t) = \mathbf{E}_{a\sim\pi(\cdot|s), s'\sim P(\cdot|s,a)}\left[r(s, a) + \gamma V_\pi(s'; t+1)\right]$$

for $0 \leq t \leq T$, with the boundary condition $V_\pi(s; T+1) := 0$. ⌟

Proof. First, for continuing tasks, we have

$$\begin{aligned}V_\pi(s) &= \mathbf{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t r(S_t, A_t)\Big| S_0 = s\right] \\ &= \mathbf{E}_\pi\left[r(S_0, A_0) + \gamma\sum_{t=1}^{\infty}\gamma^{t-1} r(S_t, A_t)\Big| S_0 = s\right] \\ &= \mathbf{E}_{a\sim\pi(\cdot|s),\, s'\sim P(\cdot|s,a)}\left[r(s, a) + \gamma\mathbf{E}_\pi\left[\sum_{t=1}^{\infty}\gamma^{t-1} r(S_t, A_t)\Big| S_0 = s, A_0 = a, S_1 = s'\right]\right] \\ &= \mathbf{E}_{a\sim\pi(\cdot|s),\, s'\sim P(\cdot|s,a)}\left[r(s, a) + \gamma\mathbf{E}_\pi\left[\sum_{t=1}^{\infty}\gamma^{t-1} r(S_t, A_t)\Big| S_1 = s'\right]\right] \\ &= \mathbf{E}_{a\sim\pi(\cdot|s),\, s'\sim P(\cdot|s,a)}\left[r(s, a) + \gamma V_\pi(s')\right].\end{aligned}$$

Note that the fourth equality follows from the Markov property: conditioned on s', the expectation is independent of $S_0 = s, A_0 = a$. Similarly, for finite-horizon episodic tasks, we have

$$\begin{aligned} V_\pi(s;t) &= \mathbf{E}_\pi\left[\sum_{k=t}^{T}\gamma^{k-t}r(S_k,A_k)\Big|S_t=s\right] \\ &= \mathbf{E}_\pi\left[r(S_t,A_t)+\gamma\sum_{k=t+1}^{T}\gamma^{k-(t+1)}r(S_k,A_k)\Big|S_t=s\right] \\ &= \mathbf{E}_{a\sim\pi(\cdot|s),\,s'\sim P(\cdot|s,a)}\left[r(s,a)+\gamma\mathbf{E}_\pi\left[\sum_{k=t+1}^{T}\gamma^{k-(t+1)}r(S_k,A_k)\Big|S_{t+1}=s'\right]\right] \\ &= \mathbf{E}_{a\sim\pi(\cdot|s),\,s'\sim P(\cdot|s,a)}\left[r(s,a)+\gamma V_\pi(s';t+1)\right]. \end{aligned}$$

Also note that $V_\pi(s;T+1)=0$ follows from the fact that the agent receives no rewards after time T. □

Proposition 1.6 (Bellman Equation for action-value function)**.** *For any policy π, any state s and action a, the Bellman equation for continuing tasks is:*

$$Q_\pi(s,a) = \mathbf{E}_\pi\left[r+\gamma V_\pi(s') \mid S_0=s, A_0=a\right].$$

For finite-horizon episodic tasks with horizon T, we have

$$Q_\pi(s,a;t) = \mathbf{E}_\pi\left[r+\gamma V_\pi(s';t+1) \mid S_t=s, A_t=a\right]$$

for $0\le t\le T$, with $V_\pi(s;T+1):=0$. ⌟

The proof of this is very similar to the last proposition.

Proposition 1.7. *We can express the state-value function in terms of the action-value function as follows:*

$$V_\pi(s) = \mathbf{E}_{a\sim\pi(\cdot|s)}\left[Q_\pi(s,a)\right].$$

For finite-horizon episodic tasks with horizon T, we have

$$V_\pi(s;t) = \mathbf{E}_{a\sim\pi(\cdot|s)}\left[Q_\pi(s,a;t)\right]$$

for $0\le t\le T$. ⌟

This follows from the definition of the state-value function and the action-value function.

1.2. **Bellman Operators**

The Bellman expectation operator, $\mathcal{T}_\pi$, for a given policy π is an operator that updates a state-value function to a new state-value function that is "closer" to the state-value function of π, V_π.

Definition 1.8 (Bellman Expectation Operator)**.** *Let $\mathcal{V}$ denote the space of real-valued functions on the state space $\mathcal{S}$, i.e. $\mathcal{V} = \{V : \mathcal{S} \to \mathbb{R}\}$. Given a stationary policy π, the Bellman expectation operator $\mathcal{T}_\pi : \mathcal{V} \to \mathcal{V}$ is defined by*

$$(\mathcal{T}_\pi V)(s) = \mathbf{E}_{a\sim\pi(\cdot|s),\, s'\sim P(\cdot|s,a)} \left[r(s,a) + \gamma V(s')\right].$$

⌟

We will now formalize that this operator sends the value function to a state-value function that is closer to V_π. The proof requires Banach fixed point theorem from topology, the discussion of which we defer to §A.

Proposition 1.9 (Contraction of the Bellman Expectation Operator)**.** *For any stationary policy π, the Bellman expectation operator $\mathcal{T}_\pi$ is a γ-contraction with respect to the supremum norm $\|V\|_\infty := \max_{s\in\mathcal{S}} |V(s)|$. In other words, for any $V, W \in \mathcal{V}$,*

$$\|\mathcal{T}_\pi V - \mathcal{T}_\pi W\|_\infty \leq \gamma \|V - W\|_\infty.$$

Furthermore, $\mathcal{T}_\pi$ admits a unique fixed point, which is the state-value function V_π, meaning

$$\mathcal{T}_\pi V_\pi = V_\pi.$$

⌟

Proof. Let $V, W \in \mathcal{V}$. For any state $s \in \mathcal{S}$, we have

$$\begin{aligned}
&|(\mathcal{T}_\pi V)(s) - (\mathcal{T}_\pi W)(s)| \\
&= \left|\mathbf{E}_{a\sim\pi(\cdot|s),\, s'\sim P(\cdot|s,a)}[\gamma(V(s') - W(s'))]\right| \\
&\leq \gamma \mathbf{E}_{a\sim\pi(\cdot|s),\, s'\sim P(\cdot|s,a)}[|V(s') - W(s')|] && \text{(Triangle inequality)} \\
&\leq \gamma \|V - W\|_\infty.
\end{aligned}$$

Since this is true for any state s, taking the maximum over $s \in \mathcal{S}$, we get

$$\|\mathcal{T}_\pi V - \mathcal{T}_\pi W\|_\infty \leq \gamma \|V - W\|_\infty.$$

Since $\gamma < 1$, this establishes that $\mathcal{T}_\pi$ is a contraction. Since $\mathcal{S}$ is finite, the space $(\mathcal{V}, \|\cdot\|_\infty)$ is complete (we prove this in §A). Therefore, by the Banach fixed-point theorem, $\mathcal{T}_\pi$ admits a unique fixed point. Now, we know that Bellman equation gives

$$V_\pi(s) = \mathbf{E}_{a\sim\pi(\cdot|s),\, (s',r)\sim P(\cdot,\cdot|s,a)}[r + \gamma V_\pi(s')].$$

Therefore, using the definition of the Bellman expectation operator:

$$\mathcal{T}_\pi V_\pi = V_\pi.$$

So, V_π is the unique fixed point of $\mathcal{T}_\pi$. □

The Bellman optimality operator updates a value function towards a different kind of state-value function. Unlike the Bellman expectation operator, $\mathcal{T}_\pi$, which udpates the value function towards V_π, the Bellman optimality operator will update it towards what we will later define to be V_*.

Definition 1.10 (Bellman Optimality Operator)**.** *The Bellman optimality operator $\mathcal{T}_* : \mathcal{V} \to \mathcal{V}$ is defined by*

$$(\mathcal{T}_* V)(s) = \max_{a\in\mathcal{A}} \mathbf{E}_{s'\sim P(\cdot|s,a)} \left[r(s,a) + \gamma V(s')\right].$$

⌟

Proposition 1.11 (Contraction of the Bellman Optimality Operator)**.** *The Bellman optimality operator $\mathcal{T}_*$ is a γ-contraction with respect to the supremum norm. In other words, for any $V, W \in \mathcal{V}$,*

$$\|\mathcal{T}_* V - \mathcal{T}_* W\|_\infty \leq \gamma \|V - W\|_\infty.$$

Furthermore, $\mathcal{T}_*$ admits a unique fixed point. ⌟

Proof. First, recall that for any two finite collections of real numbers $\{x_a\}_{a\in\mathcal{I}}$ and $\{y_a\}_{a\in\mathcal{I}}$, we have

$$\left|\max_{a\in\mathcal{I}} x_a - \max_{a\in\mathcal{I}} y_a\right| \le \max_{a\in\mathcal{I}} |x_a - y_a|.$$

Therefore, for any $V, W \in \mathcal{V}$ and $s \in \mathcal{S}$, we have

$$\begin{aligned}
&|(\mathcal{T}_* V)(s) - (\mathcal{T}_* W)(s)| \\
&= \left|\max_{a\in\mathcal{A}} \mathbf{E}_{(s',r)\sim P(\cdot,\cdot|s,a)}[r + \gamma V(s')] - \max_{a\in\mathcal{A}} \mathbf{E}_{(s',r)\sim P(\cdot,\cdot|s,a)}[r + \gamma W(s')]\right| \\
&\le \max_{a\in\mathcal{A}} \left|\mathbf{E}_{s'\sim P(\cdot|s,a)}[\gamma(V(s') - W(s'))]\right| \\
&\le \gamma \max_{a\in\mathcal{A}} \mathbf{E}_{s'\sim P(\cdot|s,a)}[|V(s') - W(s')|] \\
&\le \gamma \|V - W\|_\infty.
\end{aligned}$$

Taking the maximum over all states gives

$$\|\mathcal{T}_* V - \mathcal{T}_* W\|_\infty \le \gamma \|V - W\|_\infty.$$

This proves that $\mathcal{T}_*$ is a contraction. Again, since $(\mathcal{V}, \|\cdot\|_\infty)$ is complete and $\gamma < 1$, the Banach fixed-point theorem guarantees that $\mathcal{T}_*$ admits a unique fixed point. □

Now, we prove arguably the most important guarantee we have so far. We show that in a finite horizon MDP, there exists an optimal policy that is deterministic and stationary.

Proposition 1.12. *Consider a finite, time-homogeneous MDP with bounded rewards and discount factor $\gamma \in [0, 1)$. Then there exists a deterministic stationary policy π_* that is simultaneously optimal from every state, i.e.*

$$V_{\pi_*}(s) = \max_\pi V_\pi(s)$$

for every $s \in \mathcal{S}$, where the maximum is taken over all admissible policies. ⌟

Proof. By the preceding proposition, the Bellman optimality operator $\mathcal{T}_*$ admits a unique fixed point. Let $V_* \in \mathcal{V}$ denote this fixed point, i.e. $\mathcal{T}_* V_* = V_*$. Then, for every state s, we know that

$$V(s) = \max_{a\in\mathcal{A}} \mathbf{E}_{(s',r)\sim P(\cdot,\cdot|s,a)}[r + \gamma V_*(s')].$$

Now, for each state s, choose an action

$$a_*(s) \in \arg\max_{a\in\mathcal{A}} \mathbf{E}_{(s',r)\sim P(\cdot,\cdot|s,a)}[r + \gamma V_*(s')].$$

Such an action exists because $\mathcal{A}$ is finite. Now, we construct a deterministic stationary policy π_* by

$$\pi_*(a \mid s) = \begin{cases} 1, & a = a_*(s), \\ 0, & \text{otherwise.} \end{cases}$$

Clearly, this is deterministic and stationary. By construction, for every state s,

$$\begin{aligned}
(\mathcal{T}_{\pi_*} V_*)(s) &= \mathbf{E}_{s'\sim P(\cdot|s,a_*(s))}[r + \gamma V_*(s')] \\
&= (\mathcal{T}_* V)(s) \\
&= V(s).
\end{aligned}$$

Therefore, $\mathcal{T}_{\pi_*} V_* = V_*$. But we have already established that $\mathcal{T}_{\pi_*}$ admits a unique fixed point, which is precisely V_{π_*}. So,

$$V_* = V_{\pi_*}.$$

All that remains is for us to show that π_* is the optimal policy, i.e. $\pi_* \ge \pi$ for all π. Since $\mathcal{T}_* V_* = V_*$, we have, for every state-action pair (s, a),

$$V_*(s) \ge \mathbf{E}_{(s',r)\sim P(\cdot,\cdot|s,a)}[r + \gamma V_*(s')].$$

Consider any policy π; the policy need not be either stationary or Markovian (meaning it can condition on more than just the current state, i.e. it can condition on a history). Let H_t denote

the history observed through state S_t. By the Markov property, the last equation also implies that

$$V_*(S_t) \geq \mathbf{E}_\pi[R_t + \gamma V_*(S_{t+1}) \mid H_t, A_t].$$

Taking conditional expectations given $S_0 = s$ and applying the law of iterated expectations, we get

$$\mathbf{E}_\pi[V_*(S_t) \mid S_0 = s] \geq \mathbf{E}_\pi[R_t + \gamma V_*(S_{t+1}) \mid S_0 = s].$$

In particular, setting $t = 0$, we have

$$V_*(s) \geq \mathbf{E}_\pi[R_0 + \gamma V_*(S_1) \mid S_0 = s].$$

Now, applying the same inequality at $t = 1$, we get

$$\mathbf{E}_\pi[V_*(S_1) \mid S_0 = s] \geq \mathbf{E}_\pi[R_1 + \gamma V_*(S_2) \mid S_0 = s].$$

Substituting this into the preceding inequality yields

$$\begin{aligned} V_*(s) &\geq \mathbf{E}_\pi[R_0 + \gamma V_*(S_1) \mid S_0 = s] \\ &\geq \mathbf{E}_\pi[R_0 + \gamma R_1 + \gamma^2 V_*(S_2) \mid S_0 = s]. \end{aligned}$$

Repeating this argument, we obtain

$$\begin{aligned} V_*(s) &\geq \mathbf{E}_\pi[R_0 + \gamma V_*(S_1) \mid S_0 = s] \\ &\geq \mathbf{E}_\pi[R_0 + \gamma R_1 + \gamma^2 V_*(S_2) \mid S_0 = s] \\ &\geq \mathbf{E}_\pi[R_0 + \gamma R_1 + \gamma^2 R_2 + \gamma^3 V_*(S_3) \mid S_0 = s] \\ &\geq \cdots \\ &\geq \mathbf{E}_\pi\left[\sum_{t=0}^{n-1} \gamma^t R_t + \gamma^n V_*(S_n) \Big| S_0 = s\right]. \end{aligned}$$

Since $\mathcal{S}$ is finite, the function V_* is bounded. Hence,

$$\lim_{n\to\infty} \gamma^n \mathbf{E}_\pi[V_*(S_n) \mid S_0 = s] = 0.$$

Taking $n \to \infty$, we obtain

$$\begin{aligned} V_*(s) &\geq \mathbf{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R_t \Big| S_0 = s\right] \\ &= V_\pi(s). \end{aligned}$$

Since this inequality holds for every admissible policy π, we have

$$V_*(s) \geq \max_\pi V_\pi(s).$$

On the other hand, since $V_* = V_{\pi_*}$ and π_* is itself an admissible policy,

$$V_*(s) = V_{\pi_*}(s) \leq \max_\pi V_\pi(s).$$

Combining the two inequalities gives

$$V_{\pi_*}(s) = V_*(s) = \max_\pi V_\pi(s) = V_*(s)$$

for every $s \in \mathcal{S}$. Therefore, π_* is a deterministic stationary policy that is simultaneously optimal from every state. □

1.3. **Optimal Policy**

Definition 1.13 (Optimal Policy)**.** *With the definition of value functions, we can define an ordering over policies. We say $\pi \geq \pi'$ if and only if $V_\pi(s) \geq V_{\pi'}(s)$ for all $s \in \mathcal{S}$. The optimal policy, denoted by π_*, is the policy that satisfies $\pi_* \geq \pi$ for all π.* ⌟

Definition 1.14 (Optimal Value Functions)**.** *The optimal state-value function is the maximum expected return achievable from a given state s over all policies:*

$$V_*(s) = \max_\pi V_\pi(s).$$

Similarly, the optimal action-value function is defined as

$$Q_*(s,a) = \max_\pi Q_\pi(s,a).$$

In other words, $Q_*(s,a)$ is the expected return from taking action a in state s and acting optimally thereafter. ⌟

Proposition 1.15 (Bellman Optimality Equations)**.** *The optimal value functions satisfy*

$$V_*(s) = \max_{a \in \mathcal{A}} Q_*(s,a)$$

and

$$Q_*(s,a) = \mathbf{E}_{s' \sim P(\cdot|s,a)} \left[r + \gamma V_*(s') \right].$$

Therefore,

$$V_*(s) = \max_{a \in \mathcal{A}} \mathbf{E}_{s' \sim P(\cdot|s,a)} \left[r + \gamma V_*(s') \right].$$

⌟

Proof. First, consider continuing tasks. By the existence of an optimal stationary policy as we have already proven in Proposition 1.12, there exists π_* such that $V_{\pi_*}(s) = V_*(s)$ for every state s. Thus, for any policy π,

$$\begin{aligned} Q_\pi(s,a) &= \mathbf{E}_{s' \sim P(\cdot|s,a)}[r(s,a) + \gamma V_\pi(s')] \\ &\leq \mathbf{E}_{s' \sim P(\cdot|s,a)}[r(s,a) + \gamma V_*(s')] \\ &= Q_{\pi_*}(s,a). \end{aligned}$$

So,

$$Q_*(s,a) = Q_{\pi_*}(s,a) = \mathbf{E}_{s' \sim P(\cdot|s,a)}[r(s,a) + \gamma V_*(s')].$$

Moreover, we have already established that V_* is the unique fixed point of the Bellman optimality operator $\mathcal{T}_*$. Therefore,

$$\begin{aligned} V_*(s) &= (\mathcal{T}_* V_*)(s) && \text{(stationary point of } \mathcal{T}_*) \\ &= \max_{a \in \mathcal{A}} \mathbf{E}_{s' \sim P(\cdot|s,a)}[r(s,a) + \gamma V_*(s')] && \text{(definition of } \mathcal{T}_*) \\ &= \max_{a \in \mathcal{A}} Q_*(s,a). \end{aligned}$$

□

1.4. **State Distributions**

We will often require some notion of the distribution of states visited under a policy or the distribution of states from which trajectories begin. Let ρ denote the initial-state distribution, and let τ denote the number of decision time steps in a trajectory. For continuing tasks, we set $\tau = \infty$, whereas for episodic tasks with a finite horizon T, we have $\tau \leq T$. We also let $\mathcal{S}^+ \subseteq \mathcal{S}$ denote the set of nonterminal states, and assume that ρ is supported on $\mathcal{S}^+$.

- $\mathbb{P}_{\pi_\theta}(S_k = s,\ k < \tau \mid S_0 = s_0)$ is the probability of reaching a nonterminal state s from a fixed initial state s_0 in k time steps under policy π_θ. For $k \geq 1$, we have

$$\begin{aligned}&\mathbb{P}_{\pi_\theta}(S_k = s,\ k < \tau \mid S_0 = s_0)\\ &= \sum_{s_1,\ldots,s_{k-1} \in \mathcal{S}^+} \sum_{a_0,\ldots,a_{k-1}} \prod_{t=0}^{k-1} \pi_\theta(a_t \mid s_t) P(s_{t+1} \mid s_t, a_t),\end{aligned}$$

  where $s_k = s \in \mathcal{S}^+$. For $k = 0$, the probability is simply

$$\mathbb{P}_{\pi_\theta}(S_0 = s,\ 0 < \tau \mid S_0 = s_0) = \mathbf{1}\{s = s_0\}.$$

- When the initial state is sampled from $S_0 \sim \rho$, we define the probability of reaching state s in k time steps under policy π_θ as

$$\begin{aligned}&\mathbb{P}_{\pi_\theta}(S_k = s,\ k < \tau \mid S_0 \sim \rho)\\ &= \sum_{s_0,\ldots,s_{k-1} \in \mathcal{S}^+} \sum_{a_0,\ldots,a_{k-1}} \rho(s_0) \prod_{t=0}^{k-1} \pi_\theta(a_t \mid s_t) P(s_{t+1} \mid s_t, a_t),\end{aligned}$$

  where $s_k = s \in \mathcal{S}^+$ and $k \geq 1$. For $k = 0$, this probability is $\rho(s)$. In episodic tasks, the probability is zero for $k \geq T$. Equivalently, by the law of total probability,

$$\mathbb{P}_{\pi_\theta}(S_k = s,\ k < \tau \mid S_0 \sim \rho) = \sum_{s_0} \rho(s_0) \mathbb{P}_{\pi_\theta}(S_k = s,\ k < \tau \mid S_0 = s_0).$$

- The *unnormalized discounted state occupancy measure* is defined as

$$p^{\pi_\theta}(s) := \sum_{k=0}^{\infty} \gamma^k \mathbb{P}_{\pi_\theta}(S_k = s,\ k < \tau \mid S_0 \sim \rho).$$

  This measures the expected discounted number of visits to state s under π_θ. In general, this is not a probability distribution, since

$$\sum_{s \in \mathcal{S}} p^{\pi_\theta}(s) \neq 1.$$

- The *normalized discounted state occupancy distribution* is defined as

$$d^{\pi_\theta}(s) := \frac{p^{\pi_\theta}(s)}{Z^{\pi_\theta}},$$

  where

$$Z^{\pi_\theta} := \sum_{s \in \mathcal{S}} p^{\pi_\theta}(s)$$

  We assume that $Z^{\pi_\theta} > 0$. By construction,

$$\sum_{s \in \mathcal{S}} d^{\pi_\theta}(s) = 1.$$

  Thus, d^{π_θ} is a probability distribution over states, weighted by their discounted visitation frequencies.
- In particular, for continuing tasks with $\gamma < 1$, we have

$$Z^{\pi_\theta} = \sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma},$$

  and therefore

$$d^{\pi_\theta}(s) = (1-\gamma) p^{\pi_\theta}(s).$$

  For finite-horizon episodic tasks, however, the normalization constant generally depends on the policy and the distribution of episode lengths.

These distributions show up in useful ways in various places throughout these notes. As an example, we can rewrite the expected return in terms of the discounted state occupancy distribution as follows.

Theorem 1.16. *Let ρ be the initial-state distribution and let π_θ be a policy. Suppose either that $\gamma < 1$ or that the task is episodic with a finite horizon, and suppose that rewards are bounded. Then the expected discounted return*

$$V(\pi_\theta) := \mathbf{E}_{s_0 \sim \rho}\left[V_{\pi_\theta}(s_0)\right]$$

can be written as

$$V(\pi_\theta) = Z^{\pi_\theta} \mathbf{E}_{\substack{s \sim d^{\pi_\theta} \\ a \sim \pi_\theta(\cdot|s)}}\left[r(s,a)\right].$$

In particular, for continuing tasks with $\gamma < 1$, we have

$$V(\pi_\theta) = \frac{1}{1-\gamma} \mathbf{E}_{\substack{s \sim d^{\pi_\theta} \\ a \sim \pi_\theta(\cdot|s)}}\left[r(s,a)\right].$$

⌟

Proof. By definition of the value function,

$$\begin{aligned} V(\pi_\theta) &= \mathbf{E}_{s_0 \sim \rho}\left[V_{\pi_\theta}(s_0)\right] \\ &= \mathbf{E}_{s_0 \sim \rho}\left[\mathbf{E}_{\pi_\theta}\left[\sum_{t=0}^{\tau-1} \gamma^t r(S_t, A_t) \,\middle|\, S_0 = s_0\right]\right] \\ &= \sum_{t=0}^{\infty} \gamma^t \sum_{s \in \mathcal{S}} \mathbb{P}_{\pi_\theta}(S_t = s,\ t < \tau \mid S_0 \sim \rho) \sum_a \pi_\theta(a \mid s) r(s,a). \end{aligned}$$

Rearranging the sums gives

$$\begin{aligned} V(\pi_\theta) &= \sum_{s \in \mathcal{S}} \left(\sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{\pi_\theta}(S_t = s,\ t < \tau \mid S_0 \sim \rho)\right) \cdot \sum_a \pi_\theta(a \mid s) r(s,a) \\ &= \sum_{s \in \mathcal{S}} p^{\pi_\theta}(s) \sum_a \pi_\theta(a \mid s) r(s,a). \end{aligned}$$

Now, by the definition of the normalized discounted state occupancy distribution, we have

$$p^{\pi_\theta}(s) = Z^{\pi_\theta} d^{\pi_\theta}(s).$$

Therefore,

$$\begin{aligned} V(\pi_\theta) &= Z^{\pi_\theta} \sum_s d^{\pi_\theta}(s) \sum_a \pi_\theta(a \mid s) r(s,a) \\ &= Z^{\pi_\theta} \mathbf{E}_{\substack{s \sim d^{\pi_\theta} \\ a \sim \pi_\theta(\cdot|s)}}\left[r(s,a)\right]. \end{aligned}$$

Finally, for continuing tasks with $\gamma < 1$, we have

$$Z^{\pi_\theta} = \sum_{t=0}^{\infty} \gamma^t = \frac{1}{1-\gamma},$$

and hence

$$V(\pi_\theta) = \frac{1}{1-\gamma} \mathbf{E}_{\substack{s \sim d^{\pi_\theta} \\ a \sim \pi_\theta(\cdot|s)}}\left[r(s,a)\right].$$

□

2. Policy Gradient Methods

In policy gradient methods, we directly parametrise the policy to be $\pi_\theta(a|s)$ and our goal becomes to find π_θ that maximizes expected returns. In contrast to value-based methods where we first learn value functions and then extract a policy out of the value function, in policy gradient methods we directly learn a policy that does not consult a value function. We may still learn a value function in order to help learn our parametrized optimal policy; for example, we will find that value functions can be useful for reducing the variance in our updates. However, the value function will not be required for our action selection (in the sense that, when selecting which action to take, we will not query the value function). Methods that incorporate both a parametrized policy and a parametrized learned value function are often called *actor-critic methods*.

Policy gradient methods are useful for generating stochastic policies. A useful example of a parametrized policy to keep in mind is a Gaussian policy (especially for continuous action spaces) that is parametrized as $a \sim \mathcal{N}(\mu_\theta(s), \Sigma) =: \pi_\theta(\cdot \mid s)$ (we usually set Σ to be a diagonal matrix). The policy can be parametrized in any way as long as it is differentiable with respect to its parameters. More complex parameterizations can be used; the choice of policy parameterization is sometimes a good way to inject prior knowledge about the desired form of the policy into the reinforcement learning system. In any case, we generally want to encourage the policy to explore and, to do so, we usually require that the policy never becomes deterministic i.e. $\pi_\theta(a \mid s) \in (0, 1), \forall s, a$.

Policy gradient methods are usually suitable for continuous action spaces and for high-dimensional action spaces. However, a common phenomenon with policy gradient methods is that these policies tend to be hard to evaluate and can have a high variance.

2.1. **Policy Gradient**

Recall that our goal is to learn the policy that maximizes the expected sum of reward. As such, we define our objective to be

$$J(\theta) = \mathbf{E}_{s_0 \sim \rho}\left[V_{\pi_\theta}(s_0)\right]$$

where ρ is the initial state distribution of our MDP. We want to update our policy via gradient descent on this objective, i.e. our update rule should look like

$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$$

where α is the learning rate.

Our first goal is to compute the gradient of this objective. The policy gradient theorem provides us this:

Theorem 2.1 (Policy Gradient Theorem)**.** *Consider a finite MDP with a differentiable policy π_θ, bounded rewards, and suppose that the reward function and transition dynamics are independent of θ. Suppose either that $\gamma < 1$ or that the task is episodic with a finite horizon. Let ρ be an initial-state distribution independent of θ. Define*

$$J(\theta) := \mathbf{E}_{s_0 \sim \rho}\left[V_{\pi_\theta}(s_0)\right].$$

Then

$$\nabla_\theta J(\theta) = \sum_{s \in \mathcal{S}} p^{\pi_\theta}(s) \sum_{a \in \mathcal{A}} Q_{\pi_\theta}(s,a) \nabla_\theta \pi_\theta(a \mid s),$$

where p^{π_θ} is the unnormalized discounted state occupancy measure, defined by

$$p^{\pi_\theta}(s) := \sum_{k=0}^{\infty} \gamma^k \mathbb{P}_{\pi_\theta}(S_k = s,\ k \mid S_0 \sim \rho).$$

Furthermore, if $\pi_\theta(a \mid s) > 0$ for every state-action pair, then

$$\nabla_\theta J(\theta) = \sum_{s \in \mathcal{S}} p^{\pi_\theta}(s) \mathbf{E}_{a \sim \pi_\theta(\cdot \mid s)}\left[Q_{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(a \mid s)\right].$$

⌟

Proof. For any state s, we see that

$$\nabla_\theta V_{\pi_\theta}(s) = \nabla_\theta \left(\sum_a \pi_\theta(a \mid s) Q_{\pi_\theta}(s,a) \right) = \sum_a \nabla_\theta \pi_\theta(a \mid s) Q_{\pi_\theta}(s,a) + \sum_a \pi_\theta(a \mid s) \nabla_\theta Q_{\pi_\theta}(s,a).$$

We already saw that

$$Q_{\pi_\theta}(s,a) = r(s,a) + \gamma \sum_{s'} P(s' \mid s,a) V_{\pi_\theta}(s').$$

Since the reward function and transition probabilities do not depend on θ,

$$\nabla_\theta Q_{\pi_\theta}(s,a) = \gamma \sum_{s'} P(s' \mid s,a) \nabla_\theta V_{\pi_\theta}(s').$$

Substituting this into the last equation gives

$$\begin{aligned} \nabla_\theta V_{\pi_\theta}(s) = &\sum_a \nabla_\theta \pi_\theta(a \mid s) Q_{\pi_\theta}(s,a) \\ &+ \gamma \sum_a \sum_{s'} \pi_\theta(a \mid s) P(s' \mid s,a) \nabla_\theta V_{\pi_\theta}(s'). \end{aligned}$$

Recursively expanding, we get:

$$\nabla_\theta V_{\pi_\theta}(s) = \sum_{k=0}^{\infty} \gamma^k \sum_{x \in \mathcal{S}} \mathbb{P}_{\pi_\theta}(S_k = x,\ k < \tau \mid S_0 = s) \sum_a \nabla_\theta \pi_\theta(a \mid x) Q_{\pi_\theta}(x,a).$$

where τ denotes the number of decision time steps in the trajectory, with $\tau = \infty$ for continuing tasks. As such, note that in episodic tasks, only visits before termination are counted.

Now, we take the derivative of $J(\theta)$. Since ρ is independent of θ, we have

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \sum_{s_0} \rho(s_0) V_{\pi_\theta}(s_0) \\
&= \sum_{s_0} \rho(s_0) \nabla_\theta V_{\pi_\theta}(s_0) \\
&= \sum_{s_0} \rho(s_0) \sum_{k=0}^{\infty} \gamma^k \sum_{x \in \mathcal{S}} \mathbb{P}_{\pi_\theta}(S_k = x,\ k < \tau \mid S_0 = s_0) \sum_a \nabla_\theta \pi_\theta(a \mid x) Q_{\pi_\theta}(x, a). \\
&= \sum_{x \in \mathcal{S}} \left(\sum_{k=0}^{\infty} \gamma^k \sum_{s_0} \rho(s_0) \mathbb{P}_{\pi_\theta}(S_k = x,\ k < \tau \mid S_0 = s_0) \right) \sum_a \nabla_\theta \pi_\theta(a \mid x) Q_{\pi_\theta}(x, a). \\
&= \sum_{x \in \mathcal{S}} \left(\sum_{k=0}^{\infty} \gamma^k \mathbb{P}_{\pi_\theta}(S_k = x,\ k < \tau \mid S_0 \sim \rho) \right) \sum_a \nabla_\theta \pi_\theta(a \mid x) Q_{\pi_\theta}(x, a). \\
&= \sum_{x \in \mathcal{S}} p^{\pi_\theta}(x) \sum_a \nabla_\theta \pi_\theta(a \mid x) Q_{\pi_\theta}(x, a).
\end{aligned}$$

Here, we used the law of total probability to get

$$\sum_{s_0} \rho(s_0) \mathbb{P}_{\pi_\theta}(S_k = x,\ k < \tau \mid S_0 = s_0) = \mathbb{P}_{\pi_\theta}(S_k = x,\ k < \tau \mid S_0 \sim \rho),$$

and the definition

$$p^{\pi_\theta}(x) := \sum_{k=0}^{\infty} \gamma^k \mathbb{P}_{\pi_\theta}(S_k = x,\ k < \tau \mid S_0 \sim \rho).$$

Finally, using

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

we get

$$\begin{aligned}
\nabla_\theta J(\theta) &= \sum_s p^{\pi_\theta}(s) \sum_a \pi_\theta(a \mid s) Q_{\pi_\theta}(s, a) \nabla_\theta \log \pi_\theta(a \mid s) \\
&= \sum_s p^{\pi_\theta}(s) \mathbf{E}_{a \sim \pi_\theta(\cdot \mid s)} \left[Q_{\pi_\theta}(s, a) \nabla_\theta \log \pi_\theta(a \mid s) \right].
\end{aligned}$$

□

By the policy gradient theorem, we have that

$$\nabla_\theta J(\theta) = \sum_{s \in \mathcal{S}} p^{\pi_\theta}(s) \sum_{a \in \mathcal{A}} Q_{\pi_\theta}(s, a) \nabla_\theta \pi_\theta(a \mid s).$$

We also know that $d^{\pi_\theta}(s) := \frac{p^{\pi_\theta}(s)}{Z^{\pi_\theta}}$, which is a probability distribution. Then, our gradient becomes:

$$\begin{aligned}
\nabla_\theta J(\theta) &= \sum_{s \in \mathcal{S}} p^{\pi_\theta}(s) \sum_{a \in \mathcal{A}} Q_{\pi_\theta}(s, a) \nabla_\theta \pi_\theta(a \mid s) \\
&= Z^{\pi_\theta} \sum_{s \in \mathcal{S}} d^{\pi_\theta}(s) \sum_{a \in \mathcal{A}} Q_{\pi_\theta}(s, a) \nabla_\theta \pi_\theta(a \mid s) \\
&= Z^{\pi_\theta} \mathbf{E}_{s \sim d^{\pi_\theta}, a \sim \pi_\theta(\cdot \mid s)} \left[Q_{\pi_\theta}(s, a) \nabla_\theta \log \pi_\theta(a \mid s) \right],
\end{aligned}$$

which is a version of the policy gradient that is particularly useful. For continuing tasks, in particular, we get the nice expression:

$$\nabla_\theta J(\theta) = \frac{1}{1 - \gamma} \mathbf{E}_{s \sim d^{\pi_\theta}, a \sim \pi_\theta(\cdot \mid s)} \left[Q_{\pi_\theta}(s, a) \nabla_\theta \log \pi_\theta(a \mid s) \right].$$

2.2. **REINFORCE**

Recall that we have derived that

$$\nabla_\theta J(\theta) = Z^{\pi_\theta} \mathbf{E}_{s\sim d^{\pi_\theta}, a\sim\pi_\theta(\cdot|s)} \left[Q_{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(a \mid s) \right]. \tag{4}$$

We aim to use this formulation of the gradient in our policy update

$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta).$$

Since Z^{π_θ} is simply a constant, we can absorb this into the learning rate. Now, we aim to find practical methods for estimating the gradient:

$$\mathbf{E}_{s\sim d^{\pi_\theta}, a\sim\pi_\theta(\cdot|s)} \left[Q_{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(a \mid s) \right].$$

In this section, we study the algorithm REINFORCE. We will derive this in a very similar manner to the policy gradient theorem but we will derive a final form that is immediately practically useful. In particular, in the policy gradient theorem we derived in 2.1, the expectation was over states sampled from $p^{\pi_\theta}(\cdot)$. In most practical implementations of REINFORCE, we sample trajectories under a policy, so it is more natural to define an expectation over trajectories instead.

We first write our objective

$$J(\theta) = \mathbf{E}_{s_o\sim\rho}[V_{\pi_\theta}(s_0)] = \mathbf{E}_{\pi_\theta}[R(\tau)],$$

where $\tau = [s_0, a_0, r_0, s_1, a_1, r_1, \cdots]$ is a trajectory and

$$R(\tau) = \sum_{t=0}^{T} \gamma^t r(s_t, a_t).$$

Now, we compute the gradient of this as follows:

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \sum_\tau P_\theta(\tau) R(\tau) \\
&= \sum_\tau R(\tau) \nabla_\theta P_\theta(\tau) \\
&= \sum_\tau R(\tau) \frac{P_\theta(\tau)}{P_\theta(\tau)} \nabla_\theta P_\theta(\tau) \\
&= \sum_\tau R(\tau) \frac{P_\theta(\tau)}{P_\theta(\tau)} \nabla_\theta P_\theta(\tau) \\
&= \sum_\tau R(\tau) P_\theta(\tau) \nabla_\theta \log(P_\theta(\tau))
\end{aligned}$$

Next, we observe that

$$\begin{aligned}
\nabla_\theta \log\left(P_\theta(\tau) &= \nabla_\theta \log \left(\rho(s_0) \prod_{t=0}^{T} \pi_\theta(a_t|s_t) P(s_{t+1}|s_t, a_t) \right) \right. \\
&= \nabla_\theta \log\left(\rho(s_0)\right) + \sum_{t=0}^{T} \log \pi_\theta(a_t|s_t) + \log P(s_{t+1}|s_t, a_t) \\
&= \sum_{t=0}^{T} \nabla_\theta \log(\pi_\theta(a_t|s_t))
\end{aligned}$$

As such, our gradient becomes:

$$\nabla_\theta V(\theta) = \mathbb{E}_{\tau\sim P_\theta(\tau)} \left[\left(\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \right) \left(\sum_{t=0}^{T} \gamma^t r(s_t, a_t) \right) \right]. \tag{5}$$

Note that this is tractable as long as we can take the derivative of $\log \pi_\theta(a \mid s)$. The intuitive way to interpret this is that we are updating our policy by taking a step, in the parameter space, in the direction of increasing log probability of trajectory τ proportional to $R(\tau)$. This makes sense - our algorithm is encouraging our policy to sample trajectories that yield higher returns. On the other hand, since $\nabla_\theta \log \pi_\theta(a_t \mid s_t) = \frac{\nabla_\theta \pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)}$, the gradient norm is not only

proportional to $R(\tau)$ but also inversely proportional to the probability of sampling the trajectory. This makes sense - the algorithm provides a stronger learning signal for actions that are rare and also yield higher returns.

In almost all practical implementations, we use a Monte Carlo estimate of (5) since computing the expectation is practically infeasible in large MDPs: we sample trajectories $\tau^{(1)}, \cdots, \tau^{(N)}$ under π_θ and estimate the gradient as:

$$\widehat{\nabla_\theta J(\theta)} = \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t^{(i)} \mid s_t^{(i)}) \right) \left(\sum_{t=0}^{T} \gamma^t r(s_t^{(i)}, a_t^{(i)}) \right). \tag{6}$$

As with most Monte Carlo estimators, this gradient suffers from high variance and, very often, one needs to sample a large number of trajectories, especially when the time horizon is large, for this gradient to be reliable. We now introduce a set of techniques to diminish the variance.

2.2.1. *REINFORCE with Causality.* The first technique we introduce is based on a simple observation: an action taken at time t cannot affect rewards obtained at earlier time steps $t' < t$. Therefore, when maximizing the log probability of taking a particular action a_t from state s_t, we should only weigh it by the rewards obtained from time t onward, rather than the return of the entire trajectory. In fact, in the policy gradient theorem (see Section 2.1), the gradient $\nabla_\theta \log \pi_\theta(a \mid s)$ is multiplied by $Q_{\pi_\theta}(s, a)$, which measures the expected return from taking action a in state s and following the policy thereafter. We now formalize this intuition.

Recall that our objective is

$$J(\theta) = \mathbf{E}_{\tau \sim p_\theta} \left[\sum_{t=0}^{T} \gamma^t r_t \right].$$

For the moment, suppose we instead consider the expected reward at a particular timestep t,

$$V_t(\theta) := \mathbf{E}_{\tau \sim p_\theta}[r_t].$$

By causality, the reward r_t depends only on the trajectory up to timestep t. Thus, we can equivalently write

$$V_t(\theta) = \mathbf{E}_{\tau_{0:t} \sim p_\theta}[r_t],$$

where $\tau_{0:t} = (s_0, a_0, \ldots, s_t, a_t)$ denotes the trajectory prefix up to timestep t. Applying the log-derivative trick, we obtain

$$\begin{aligned} \nabla_\theta V_t(\theta) &= \mathbf{E}_{\tau_{0:t} \sim p_\theta} \left[r_t \nabla_\theta \log p_\theta(\tau_{0:t}) \right] \\ &= \mathbf{E}_{\tau_{0:t} \sim p_\theta} \left[r_t \sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta(a_{t'} \mid s_{t'}) \right], \end{aligned}$$

where the second equality follows because the initial-state distribution and environment dynamics do not depend on θ.

Now, observe that

$$\begin{aligned} \sum_{t=0}^{T} \gamma^t V_t(\theta) &= \sum_{t=0}^{T} \gamma^t \mathbf{E}_{\tau_{0:t} \sim p_\theta}[r_t] \\ &= \mathbf{E}_{\tau \sim p_\theta} \left[\sum_{t=0}^{T} \gamma^t r_t \right] \\ &= J(\theta). \end{aligned}$$

Therefore, we have

$$\begin{aligned}
\nabla_\theta J(\theta) &= \sum_{t=0}^{T} \gamma^t \nabla_\theta V_t(\theta) \\
&= \sum_{t=0}^{T} \gamma^t \mathbf{E}_{\tau \sim p_\theta} \left[r_t \sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta(a_{t'} \mid s_{t'}) \right] \\
&= \mathbf{E}_{\tau \sim p_\theta} \left[\sum_{t=0}^{T} \gamma^t r_t \sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta(a_{t'} \mid s_{t'}) \right] \\
&= \mathbf{E}_{\tau \sim p_\theta} \left[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \sum_{t'=t}^{T} \gamma^{t'} r_{t'} \right] \\
&= \mathbf{E}_{\tau \sim p_\theta} \left[\sum_{t=0}^{T} \gamma^t \nabla_\theta \log \pi_\theta(a_t \mid s_t) \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'} \right].
\end{aligned}$$

In the final equality, we simply exchanged the order of summation.

Now, recall that the return from time t onward is defined as

$$G_t := \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}.$$

Therefore, our gradient becomes:

$$\nabla_\theta J(\theta) = \mathbf{E}_{\tau \sim p_\theta} \left[\sum_{t=0}^{T} \gamma^t G_t \nabla_\theta \log \pi_\theta(a_t \mid s_t) \right]. \tag{7}$$

This gives us the REINFORCE gradient with causality, where each action is weighted only by the discounted return from that timestep onward. We can estimate this with a Monte Carlo estimator as well: we sample trajectories $\tau^{(1)}, \cdots, \tau^{(N)}$ under π_θ and estimate the gradient as:

$$\widehat{\nabla_\theta J(\theta)} = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t^{(i)} \mid s_t^{(i)}) \gamma^t G_t^{(i)}, \tag{8}$$

where

$$G_t^{(i)} = \sum_{t'=t}^{T} \gamma^{t'-t} r(s_{t'}^{(i)}, a_{t'}^{(i)}).$$

REINFORCE with Causality

Initialize policy parameters θ
for iteration $= 1, 2, \ldots$ **do**
  Collect N trajectories $\{\tau^i\}_{i=1}^N$ under π_θ
  for each trajectory τ^i **do**
    for each timestep t in τ^i **do**
      Compute return: $G_t^i = \sum_{t'=t}^{T} \gamma^{t'-t} r(s_{t'}^i, a_{t'}^i)$
    end for
  end for
  Update the policy parameters:

$$\widehat{\nabla_\theta J(\theta)} = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T} \gamma^t G_t^i \nabla_\theta \log \pi_\theta(a_t^i \mid s_t^i)$$

$$\theta \leftarrow \theta + \alpha \widehat{\nabla_\theta J(\theta)}$$

end for

2.2.2. *REINFORCE with Baseline.* The next technique we introduce is the notion of a baseline. We use the following gradient estimator:

$$\widehat{\nabla_\theta J(\theta)} = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T}\gamma^t \nabla_\theta \log \pi_\theta(a_t^{(i)} \mid s_t^{(i)}) \left(G_t^{(i)} - b_t(s_t^{(i)})\right), \tag{9}$$

where

$$G_t^{(i)} = \sum_{t'=t}^{T}\gamma^{t'-t} r(s_{t'}^{(i)}, a_{t'}^{(i)})$$

and $b_t(s)$ is a baseline that may depend on the state and timestep, but not on the action. We will show that subtracting such a baseline does not introduce bias into the gradient estimator. Furthermore, an appropriately chosen baseline can reduce its variance.

First, observe that for any state s,

$$\begin{aligned}
&\mathbf{E}_{a\sim\pi_\theta(\cdot|s)}\left[b_t(s)\nabla_\theta \log \pi_\theta(a \mid s)\right] \\
&= b_t(s)\sum_a \pi_\theta(a \mid s)\nabla_\theta \log \pi_\theta(a \mid s) \\
&= b_t(s)\sum_a \nabla_\theta \pi_\theta(a \mid s) \\
&= b_t(s)\nabla_\theta \sum_a \pi_\theta(a \mid s) \\
&= b_t(s)\nabla_\theta 1 \\
&= 0.
\end{aligned}$$

Therefore,

$$\mathbf{E}_{\tau\sim P_\theta}\left[\sum_{t=0}^{T}\gamma^t b_t(S_t)\nabla_\theta \log \pi_\theta(A_t \mid S_t)\right] = 0.$$

As such, we have

$$\nabla_\theta J(\theta) = \mathbf{E}_{\tau\sim P_\theta}\left[\sum_{t=0}^{T}\gamma^t (G_t - b_t(S_t))\nabla_\theta \log \pi_\theta(A_t \mid S_t)\right].$$

Thus, subtracting any state-dependent baseline does not change the expected policy gradient.

The next question we ask is: what baseline should we choose? In particular, which baselines help us reduce variance? Some common choices of baselines include

$$b := b_t(s) = \frac{1}{N}\sum_{i=1}^{N} R(\tau^{(i)}), \qquad \text{for any } s \in \mathcal{S} \text{ and any time step } t.$$

Another common choice is the baseline:

$$b_t(s) = V_{\pi_\theta,t}(s) = \mathbf{E}_{\pi_\theta}[G_t \mid S_t = s],$$

where one estimate $V_{\pi_\theta,t}(s)$ through some learned value function, which gives rise to actor critic methods (which we will discuss in the next section).

REINFORCE with baseline

Initialize policy parameters θ
for iteration $= 1, 2, \ldots$ **do**
  Collect N trajectories $\{\tau^i\}_{i=1}^N$ under π_θ
  for each trajectory τ^i **do**
    for each timestep t in τ^i **do**
      Compute return: $G_t^i = \sum_{t'=t}^T \gamma^{t'-t} r(s_{t'}^i, a_{t'}^i)$
    end for
  end for
  Compute baseline $b = \frac{1}{N} \sum_{i=1}^N R(\tau^i)$
  Update the policy parameters:

$$\widehat{\nabla_\theta J(\theta)} = \frac{1}{N} \sum_{i=1}^N \sum_{t=0}^T \gamma^t (G_t^i - b) \nabla_\theta \log \pi_\theta(a_t^i \mid s_t^i)$$

$$\theta \leftarrow \theta + \alpha \widehat{\nabla_\theta J(\theta)}$$

end for

We now discuss theoretical justifications for this choice of baseline. First, we derive the optimal baseline.

Proposition 2.2 (Optimal Baseline for REINFORCE)**.** *Fix a timestep t and a state s, and let $g_t := \nabla_\theta \log \pi_\theta(A_t \mid s)$. Suppose that the relevant second moments are finite and that $\mathbf{E}[\|g_t\|^2 \mid S_t = s] > 0$. Then, the baseline minimizing the conditional variance of $g_t(G_t - b_t(s))$ is*

$$b_t^*(s) = \frac{\mathbf{E}[\|g_t\|^2 G_t \mid S_t = s]}{\mathbf{E}[\|g_t\|^2 \mid S_t = s]}.$$

In particular, this baseline satisfies

$$\operatorname{Var}\left(g_t(G_t - b_t^*(s)) \mid S_t = s\right) \leq \operatorname{Var}\left(g_t G_t \mid S_t = s\right).$$

⌟

Proof. Recall that the variance of a random vector X is defined as

$$\operatorname{Var}(X) := \mathbf{E}[\|X - \mathbf{E}[X]\|^2] = \mathbf{E}[\|X\|^2] - \|\mathbf{E}[X]\|^2.$$

We have already shown that

$$\mathbf{E}[g_t \mid S_t = s] = \sum_a \pi_\theta(a \mid s) \nabla_\theta \log \pi_\theta(a \mid s) = \sum_a \nabla_\theta \pi_\theta(a \mid s) = \nabla_\theta \sum_a \pi_\theta(a \mid s) = 0.$$

So, for any baseline $b_t(s)$,

$$\begin{aligned}
&\mathbf{E}[g_t(G_t - b_t(s)) \mid S_t = s] \\
&= \mathbf{E}[g_t G_t \mid S_t = s] - b_t(s) \mathbf{E}[g_t \mid S_t = s] \\
&= \mathbf{E}[g_t G_t \mid S_t = s].
\end{aligned}$$

Since the expectation of $g_t(G_t - b_t(s))$ does not depend on $b_t(s)$ (as we have just shown above), minimizing the variance of $\mathbf{E}[g_t G_t \mid S_t = s]$ is equivalent to solving the following optimization problem:

$$\min_{b_t(s)} \mathbf{E}\left[\|g_t\|^2 (G_t - b_t(s))^2 \mid S_t = s\right].$$

Expanding the square gives us

$$\begin{aligned}
\mathbf{E}[\|g_t\|^2 (G_t - b_t(s))^2 \mid S_t = s] = \mathbf{E}[\|g_t\|^2 G_t^2 \mid S_t = s] \\
- 2 b_t(s) \mathbf{E}[\|g_t\|^2 G_t \mid S_t = s] \\
+ b_t(s)^2 \mathbf{E}[\|g_t\|^2 \mid S_t = s].
\end{aligned}$$

Differentiating this optimization objective with respect to $b_t(s)$ and setting the derivative equal to zero gives

$$0 = -2\mathbf{E}[\|g_t\|^2 G_t \mid S_t = s] + 2b_t(s)\mathbf{E}[\|g_t\|^2 \mid S_t = s].$$

Therefore,

$$b_t^*(s) = \frac{\mathbf{E}[\|g_t\|^2 G_t \mid S_t = s]}{\mathbf{E}[\|g_t\|^2 \mid S_t = s]}.$$

Since

$$\frac{\partial^2}{\partial b_t(s)^2}\mathbf{E}[\|g_t\|^2 (G_t - b_t(s))^2 \mid S_t = s] = 2\mathbf{E}[\|g_t\|^2 \mid S_t = s] > 0,$$

this is the unique minimizer. Since $b_t^*(s)$ minimizes the conditional variance,

$$\operatorname{Var}\left(g_t(G_t - b_t^*(s)) \mid S_t = s\right) \leq \operatorname{Var}\left(g_t G_t \mid S_t = s\right),$$

where the right-hand side corresponds to using no baseline (i.e. REINFORCE with Causality only). □

While this is the optimal baseline for REINFORCE, in practice, we do not use it, mainly because computing $\|g_t\|^2$ can be extraordinarily expensive. Instead, we often use the *value function* as the baseline, which we provide (some) justification for now:

Proposition 2.3. *Under the assumptions of the preceding proposition, suppose that $\|g_t\|^2$ and G_t are conditionally uncorrelated given $S_t = s$, i.e.,*

$$\operatorname{Cov}\left(\|g_t\|^2, G_t \mid S_t = s\right) = 0.$$

Then the optimal baseline is the state-value function:

$$b_t^*(s) = V_{\pi_\theta, t}(s).$$

In particular, the conclusion holds if $\|g_t\|^2$ and G_t are conditionally independent given $S_t = s$. ⌟

Proof. We have already proven that the optimal baseline is

$$b_t^*(s) = \frac{\mathbf{E}[\|g_t\|^2 G_t \mid S_t = s]}{\mathbf{E}[\|g_t\|^2 \mid S_t = s]}.$$

By the definition of conditional covariance, we have

$$\begin{aligned}&\operatorname{Cov}(\|g_t\|^2, G_t \mid S_t = s)\\ &= \mathbf{E}[\|g_t\|^2 G_t \mid S_t = s] - \mathbf{E}[\|g_t\|^2 \mid S_t = s]\mathbf{E}[G_t \mid S_t = s].\end{aligned}$$

Now, assume that the conditional covariance is zero. Therefore,

$$\mathbf{E}[\|g_t\|^2 G_t \mid S_t = s] = \mathbf{E}[\|g_t\|^2 \mid S_t = s]\mathbf{E}[G_t \mid S_t = s].$$

Substituting this into the expression for the optimal baseline gives

$$\begin{aligned}b_t^*(s) &= \frac{\mathbf{E}[\|g_t\|^2 \mid S_t = s]\mathbf{E}[G_t \mid S_t = s]}{\mathbf{E}[\|g_t\|^2 \mid S_t = s]}\\ &= \mathbf{E}[G_t \mid S_t = s]\\ &= V_{\pi_\theta, t}(s).\end{aligned}$$

Finally, conditional independence implies zero conditional covariance, so the conclusion also holds under the stronger assumption of conditional independence. □

This result requires a particularly strong assumption which does not hold in most practical settings. However, empirically, we tend to find this baseline in most settings to be strong.

2.3. **On-Policy Actor-Critic Methods**

Let us consider the setting of REINFORCE with baseline. In particular, suppose that we choose the baseline to be the true state-value function under the current policy, i.e. $b_t(s) = V_{\pi_\theta,t}(s)$. We use the time step dependent notation because we are considering finite-horizon episodic tasks. In the continuing tasks setting, we would replace each $V_{\pi_\theta,t}(s)$ with $V_{\pi_\theta}(s)$.

Recall that REINFORCE with baseline gives us the policy gradient

$$\nabla_\theta J(\theta) = \mathbf{E}_{\tau\sim P_\theta}\left[\sum_{t=0}^{T}\gamma^t \nabla_\theta \log \pi_\theta(A_t \mid S_t)\left(G_t - V_{\pi_\theta,t}(S_t)\right)\right]. \tag{10}$$

Now, note that the action-value function satisfies

$$Q_{\pi_\theta,t}(s,a) = \mathbf{E}_{\pi_\theta}[G_t \mid S_t = s, A_t = a].$$

As such, by the law of iterated expectations, we may replace the reward-to-go G_t with the action-value function $Q_{\pi_\theta,t}(S_t, A_t)$ without changing the expected policy gradient. We therefore obtain

$$\nabla_\theta J(\theta) = \mathbf{E}_{\tau\sim P_\theta}\left[\sum_{t=0}^{T}\gamma^t \nabla_\theta \log \pi_\theta(A_t \mid S_t) A_{\pi_\theta,t}(S_t, A_t)\right], \tag{11}$$

where

$$\begin{aligned} A_{\pi_\theta,t}(s,a) &= Q_{\pi_\theta,t}(s,a) - V_{\pi_\theta,t}(s) \\ &= r(s,a) + \gamma \mathbf{E}_{s'\sim P(\cdot|s,a)}[V_{\pi_\theta,t+1}(s')] - V_{\pi_\theta,t}(s). \end{aligned}$$

In practice, however, we do not have access to the true state-value function. Instead, actor-critic methods use a learned value function $V_{\phi,t}(s) \approx V_{\pi_\theta,t}(s)$, parameterized by ϕ. Given a transition (s_t, a_t, r_t, s_{t+1}), we estimate the advantage using

$$\delta_t = r_t + \gamma V_{\phi,t+1}(s_{t+1}) - V_{\phi,t}(s_t). \tag{12}$$

This quantity is called the *temporal-difference error*, or TD error. Here, we set $V_{\phi,t+1}(s_{t+1}) = 0$ whenever s_{t+1} is terminal or $t = T$.

Observe that if the learned value function is exact, i.e. $V_{\phi,t} = V_{\pi_\theta,t}$, then

$$\mathbf{E}[\delta_t \mid S_t = s, A_t = a] = A_{\pi_\theta,t}(s,a).$$

Thus, the TD error provides a single-sample estimate of the advantage function. This motivates the following policy gradient estimator:

$$\widehat{\nabla_\theta J(\theta)} = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T_i}\gamma^t \delta_t^{(i)} \nabla_\theta \log \pi_\theta(a_t^{(i)} \mid s_t^{(i)}), \tag{13}$$

where $T_i \le T$ denotes the final decision timestep of trajectory i.

The next question is how to train the critic $V_{\phi,t}$. One approach is to train the critic through supervised regression using Monte Carlo returns. Suppose that our training set consists of N trajectories sampled under the current policy π_θ. For each trajectory, we compute

$$G_t^{(i)} = \sum_{t'=t}^{T_i}\gamma^{t'-t} r(s_{t'}^{(i)}, a_{t'}^{(i)}).$$

We then minimize the following squared-error loss:

$$\mathcal{L}_{\mathrm{MC}}(\phi) = \frac{1}{2}\sum_{i=1}^{N}\sum_{t=0}^{T_i}\left(V_{\phi,t}(s_t^{(i)}) - G_t^{(i)}\right)^2. \tag{14}$$

Differentiating with respect to the critic parameters ϕ gives

$$\nabla_\phi \mathcal{L}_{\mathrm{MC}}(\phi) = \sum_{i=1}^{N}\sum_{t=0}^{T_i}\left(V_{\phi,t}(s_t^{(i)}) - G_t^{(i)}\right)\nabla_\phi V_{\phi,t}(s_t^{(i)}). \tag{15}$$

Thus, we can update the critic parameters using gradient descent:

$$\phi \leftarrow \phi - \alpha^\phi \nabla_\phi \mathcal{L}_{\mathrm{MC}}(\phi).$$

Alternatively, we can train the critic using a bootstrapped estimate of the target. In particular, given a transition (s_t, a_t, r_t, s_{t+1}), we define the one-step TD target

$$y_t = r_t + \gamma V_{\phi,t+1}(s_{t+1}).$$

We then consider the squared-error loss

$$\mathcal{L}_{\mathrm{TD}}(\phi) = \frac{1}{2} \sum_{i=1}^{N} \sum_{t=0}^{T_i} \left(V_{\phi,t}(s_t^{(i)}) - \mathrm{sg}(y_t^{(i)}) \right)^2, \tag{16}$$

where sg denotes the stop-gradient operation, which treats its argument as constant during differentiation. Differentiating with respect to ϕ gives

$$\nabla_\phi \mathcal{L}_{\mathrm{TD}}(\phi) = \sum_{i=1}^{N} \sum_{t=0}^{T_i} \left(V_{\phi,t}(s_t^{(i)}) - y_t^{(i)} \right) \nabla_\phi V_{\phi,t}(s_t^{(i)}) \tag{17}$$

$$= - \sum_{i=1}^{N} \sum_{t=0}^{T_i} \delta_t^{(i)} \nabla_\phi V_{\phi,t}(s_t^{(i)}). \tag{18}$$

Consequently, gradient descent on the one-step TD loss gives the critic update

$$\phi \leftarrow \phi + \alpha^\phi \delta_t \nabla_\phi V_{\phi,t}(s_t). \tag{19}$$

Combining the critic update in (19) with the policy gradient estimator in (13) gives the following one-step actor-critic algorithm.

One-Step On-Policy Actor-Critic

Input: Policy $\pi_\theta(a \mid s)$ and critic $V_{\phi,t}(s)$
Input: Learning rates $\alpha^\theta, \alpha^\phi > 0$
Initialize policy parameters θ and critic parameters ϕ
for iteration $= 1, 2, \ldots$ **do**
  Collect N trajectories $\{\tau^i\}_{i=1}^N$ under π_θ
  for each trajectory τ^i **do**
    for each timestep t in τ^i **do**
      Compute TD error: $\delta_t^i = r_t^i + \gamma V_{\phi,t+1}(s_{t+1}^i) - V_{\phi,t}(s_t^i)$,
      (Set $V_{\phi,t+1}(s_{t+1}^i) = 0$ if s_{t+1}^i is terminal or $t = T$.)
    end for
  end for
  Compute critic loss gradient:
  $$\widehat{\nabla_\phi \mathcal{L}_{\mathrm{TD}}(\phi)} = - \sum_{i=1}^{N} \sum_{t=0}^{T_i} \delta_t^i \nabla_\phi V_{\phi,t}(s_t^i)$$
  Compute policy gradient:
  $$\widehat{\nabla_\theta J(\theta)} = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T_i} \gamma^t \delta_t^i \nabla_\theta \log \pi_\theta(a_t^i \mid s_t^i)$$
  Update critic parameters:
  $$\phi \leftarrow \phi - \alpha^\phi \widehat{\nabla_\phi \mathcal{L}_{\mathrm{TD}}(\phi)}$$
  Update policy parameters:
  $$\theta \leftarrow \theta + \alpha^\theta \widehat{\nabla_\theta J(\theta)}$$
end for

2.4. **Performance Difference Lemma**

We first recall the definitions of various state distributions in §1.4. Let ρ denote the initial-state distribution, and let τ denote the number of decision time steps in a trajectory. For continuing tasks, we set $\tau = \infty$, whereas for episodic tasks with a finite horizon T, we have $\tau \leq T$. We also let $\mathcal{S}^+ \subseteq \mathcal{S}$ denote the set of nonterminal states, and assume that ρ is supported on $\mathcal{S}^+$. Throughout this section, given a trajectory $\tau = (s_0, a_0, r_0, s_1, a_1, ...)$, we define the probability of sampling it under policy π_θ to be

$$P_\theta(\tau) = \rho(s_0) \prod_{i=0}^{\infty} \pi_\theta(a_i|s_i) P(s_{i+1}|s_i, a_i).$$

We define the probability of reaching state s in k time steps under policy π_θ as

$$\begin{aligned} &\mathbb{P}_{\pi_\theta}(S_k = s \mid S_0 \sim \rho) \\ &= \sum_{s_0,\ldots,s_{k-1} \in \mathcal{S}^+} \sum_{a_0,\ldots,a_{k-1}} \rho(s_0) \prod_{t=0}^{k-1} \pi_\theta(a_t \mid s_t) P(s_{t+1} \mid s_t, a_t), \end{aligned}$$

where $s_k = s \in \mathcal{S}^+$ and $k \geq 1$. For $k = 0$, this probability is $\rho(s)$. Equivalently, by the law of total probability,

$$\mathbb{P}_{\pi_\theta}(S_k = s \mid S_0 \sim \rho) = \sum_{s_0} \rho(s_0) \mathbb{P}_{\pi_\theta}(S_k = s \mid S_0 = s_0).$$

The *unnormalized discounted state occupancy measure* is defined as

$$p^{\pi_\theta}(s) := \sum_{k=0}^{\infty} \gamma^k \mathbb{P}_{\pi_\theta}(S_k = s \mid S_0 \sim \rho).$$

This measures the expected discounted number of visits to state s under π_θ. In general, this is not a probability distribution, since

$$\sum_{s \in \mathcal{S}} p^{\pi_\theta}(s) \neq 1.$$

The *normalized discounted state occupancy distribution* is defined as

$$d^{\pi_\theta}(s) := \frac{p^{\pi_\theta}(s)}{Z^{\pi_\theta}}, \qquad \text{where } Z^{\pi_\theta} := \sum_{s \in \mathcal{S}} p^{\pi_\theta}(s).$$

For continuing tasks with $\gamma < 1$, we have

$$d^{\pi_\theta}(s) = (1 - \gamma) p^{\pi_\theta}(s).$$

Throughout this section, we will assume that $\gamma \in [0, 1)$ and that the reward function is bounded, i.e. $r(s, a) \in [R_l, R_u]$ for some finite $R_l, R_u \in \mathbb{R}$. Furthermore, we will use the notation $V(\pi) = \mathbf{E}_{\tau \sim P^\pi} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$

Lemma 2.4. *Let $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be a bounded function. Then, for any policy π_θ, we have*

$$\mathbb{E}_{\tau \sim P_\theta} \left[\sum_{t=0}^{\infty} \gamma^t f(s_t, a_t) \right] = \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d^{\pi_\theta}} \left[\mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[f(s, a) \right] \right].$$

⌟

Proof. Since f is bounded and $\gamma < 1$, we can interchange the infinite sum and expectation. Thus,

$$\begin{aligned} \mathbb{E}_{\tau \sim P_\theta} \left[\sum_{t=0}^{\infty} \gamma^t f(s_t, a_t) \right] &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\tau \sim P_\theta} \left[f(s_t, a_t) \right] \\ &= \sum_{t=0}^{\infty} \gamma^t \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mathbb{P}_{\pi_\theta}(S_t = s \mid S_0 \sim \rho) \pi_\theta(a \mid s) f(s, a) \\ &= \sum_{s \in \mathcal{S}} \left(\sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{\pi_\theta}(S_t = s \mid S_0 \sim \rho) \right) \sum_{a \in \mathcal{A}} \pi_\theta(a \mid s) f(s, a) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{s\in\mathcal{S}} p^{\pi_\theta}(s)\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[f(s,a)\right] \\
&= \frac{1}{1-\gamma}\sum_{s\in\mathcal{S}} d^{\pi_\theta}(s)\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[f(s,a)\right] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_\theta}}\left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[f(s,a)\right]\right].
\end{aligned}
$$

□

Theorem 2.5 (Performance Difference Lemma)**.** *Consider a discounted continuing task with bounded rewards and discount factor $\gamma \in [0,1)$. Given two policies π and π', for any fixed initial state $s_0 \in \mathcal{S}$, we have*

$$
V_{\pi'}(s_0) - V_\pi(s_0) = \mathbb{E}_{\tau\sim P^{\pi'}}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t) \mid s_0\right] = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi'}_{s_0}}\left[\mathbb{E}_{a\sim\pi'(\cdot|s)}\left[A_\pi(s,a)\right]\right],
$$

where the advantage function is $A_\pi(s,a) := Q_\pi(s,a) - V_\pi(s)$ and $d^{\pi'}_{s_0}$ is the discounted state distribution under π' starting from s_0:

$$
d^{\pi}_{s_0}(s) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}_\pi(S_t = s \mid S_0 = s_0).
$$

⌟

Proof. The advantage function decomposes as:

$$
\begin{aligned}
A_\pi(s,a) &= Q_\pi(s,a) - V_\pi(s) \\
&= \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a) + \gamma V_\pi(s') - V_\pi(s)\right].
\end{aligned}
$$

Now, we expand as follows:

$$
\mathbb{E}_{\tau\sim P^{\pi'}}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t) \mid s_0\right] = \mathbb{E}_{\tau\sim P^{\pi'}}\left[\sum_{t=0}^{\infty}\gamma^t\left(r(s_t,a_t) + \gamma V_\pi(s_{t+1}) - V_\pi(s_t)\right) \mid s_0\right]
$$

Note that the telescoping sum give us

$$
\mathbb{E}_{\tau\sim P^{\pi'}}\left[\sum_{t=0}^{\infty}\gamma^{t+1}V_\pi(s_{t+1}) - \gamma^t V_\pi(s_t) \mid s_0\right] = \mathbf{E}_{\tau\sim P^{\pi'}}\left[-V_\pi(s_0) \mid s_0\right].
$$

So,

$$
\begin{aligned}
\mathbb{E}_{\tau\sim P^{\pi'}}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t) \mid s_0\right] &= \mathbb{E}_{\tau\sim P^{\pi'}}\left[\sum_{t=0}^{\infty}\gamma^t\left(r(s_t,a_t) + \gamma V_\pi(s_{t+1}) - V_\pi(s_t)\right) \mid s_0\right] \\
&= \mathbb{E}_{\tau\sim P^{\pi'}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t) - V_\pi(s_0) \mid s_0\right] \\
&= V_{\pi'}(s_0) - V_\pi(s_0),
\end{aligned}
$$

Using the previous lemma, we have

$$
\mathbb{E}_{\tau\sim P^{\pi'}}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t) \mid s_0\right] = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi'}_{s_0}}\left[\mathbb{E}_{a\sim\pi'(\cdot|s)}\left[A_\pi(s,a)\right]\right].
$$

□

If we define the performance of a policy π by $V(\pi) := \mathbb{E}_{s_0\sim\rho}[V_\pi(s_0)]$, then averaging the Performance Difference Lemma over the initial-state distribution ρ gives

$$
V(\pi') - V(\pi) = \mathbb{E}_{\tau\sim P^{\pi'}}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\right] = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi'}}\left[\mathbb{E}_{a\sim\pi'(\cdot|s)}\left[A_\pi(s,a)\right]\right],
$$

where trajectories are initialized according to $s_0 \sim \rho$, and $d^{\pi'}$ is the corresponding discounted state distribution. This is a version of the performance difference lemma presented in several works such as in [SLM+17].

References

[HZAL18] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, 2018.

[KNL21] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning, 2021.

[LP09] David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, second edition, 2009.

[MR26] Martin Marek and Max Ryabinin. Score centering stabilizes off-policy reinforcement learning, 2026.

[PS08] Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomput.*, 71(7–9):1180–1190, March 2008.

[SB18] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[SLH+14] David Silver, Guy Lever, Nicolas Manfred Otto Heess, Thomas Degris, Daan Wierstra, and Martin A. Riedmiller. Deterministic policy gradient algorithms. In *International Conference on Machine Learning*, 2014.

[SLM+17] John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization, 2017.

[SWD+17] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

[Wen18] Lilian Weng. A (long) peek into reinforcement learning. *lilianweng.github.io*, 2018.

[ZDY+25] Chujie Zheng, Kai Dang, Bowen Yu, Mingze Li, Huiqiang Jiang, Junrong Lin, Yuqiong Liu, Hao Lin, Chencan Wu, Feng Hu, An Yang, Jingren Zhou, and Junyang Lin. Stabilizing reinforcement learning with llms: Formulation and practices, 2025.

Appendix A. Banach Fixed-Point Theorem

A metric space is a set X with a function $d : X \times X \to \mathbb{R}$ (called a matric) satisfying:

(1) $d(x, y) \geq 0$ for all x, y
(2) $d(x, y) = 0$ if and only if $x = y$
(3) $d(x, y) = d(y, x)$ for all xy
(4) $d(x, y) \leq d(x, z) + d(z, y)$ for all $x, y, z \in X$.

Definition A.1 (Cauchy Sequence)**.** *Let (X, d) be a metric space. A sequence $(x_n)_{n\geq 0}$ in X is called a Cauchy sequence if, for every $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that*

$$d(x_n, x_m) < \varepsilon$$

for all $n, m \geq N$. ⌟

Lemma A.2. *If the sequence (x_n) converges to some x, then the sequence is Cauchy.* ⌟

Proof. Given the sequence converges, for any $\epsilon > 0$, there exists N such that for all $n > n$, $d(x_n, x) < \epsilon/2$. Then, for all $n, m > n$, we have that $d(x_n, x_m) \leq d(x_n, x) + d(x_m, x) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$. □

The converse of this lemma is not true in general but is true in a special class of spaces called complete metric spaces.

Definition A.3 (Complete Metric Space)**.** *A metric space (X, d) is said to be complete if every Cauchy sequence in X converges to some element of X.* ⌟

Definition A.4 (Contraction)**.** *Let (X, d) be a metric space. A map $\phi : X \to X$ is called a contraction if there exists a constant $c \in [0, 1)$ such that*

$$d(\phi(x), \phi(y)) \leq c\, d(x, y), \qquad \forall x, y \in X.$$

A point $x \in X$ is called a fixed point of ϕ if $\phi(x) = x$. ⌟

Proposition A.5 (Banach Fixed-Point Theorem)**.** *Let (X, d) be a nonempty complete metric space and let $\phi : X \to X$ be a contraction with the contraction constant being $c \in [0, 1)$. Then ϕ admits a unique fixed point, i.e. there exists a unique $x^* \in X$ such that*

$$\phi(x^*) = x^*.$$

Furthermore, for any $x_0 \in X$, the sequence $(x_n)_{n\geq 0}$ defined recursively by $x_{n+1} = \phi(x_n)$ converges to x^*, i.e.

$$\lim_{n\to\infty} \phi^n(x_0) = x^*.$$

In particular, the convergence satisfies

$$d(x_n, x^*) \leq c^n d(x_0, x^*).$$

⌟

Proof. Pick any $x_0 \in X$ and define a sequence $(x_n)_{n\geq 0}$ recursively by $x_{n+1} = \phi(x_n)$. We claim that this sequence is Cauchy: since ϕ is a contraction, we know that

$$\begin{aligned} d(x_n, x_{n+1}) &= d(\phi(x_{n-1}), \phi(x_n)) \\ &\leq c\, d(x_{n-1}, x_n). \end{aligned}$$

Then, under recursion, we get:

$$\begin{aligned} d(x_n, x_{n+1}) &\leq c\, d(x_{n-1}, x_n) \\ &\leq c^2 d(x_{n-2}, x_{n-1}) \\ &\leq \cdots \\ &\leq c^n d(x_0, x_1). \end{aligned}$$

Now, for any $n < m$, since d is a metric that satisfies the triangle inequality,

$$\begin{aligned} d(x_n, x_m) &\leq \sum_{i=n}^{m-1} d(x_i, x_{i+1}) \\ &\leq \sum_{i=n}^{m-1} c^i d(x_0, x_1) \\ &= c^n(1 + c + \cdots + c^{m-n-1}) d(x_0, x_1) \\ &\leq c^n \left(\sum_{i=0}^{\infty} c^i \right) d(x_0, x_1) \\ &= \frac{c^n}{1-c} d(x_0, x_1). \end{aligned}$$

We defined $c < 1$, so the right-hand side converges to 0 as $n \to \infty$. Therefore, $(x_n)_{n \geq 0}$ is a Cauchy sequence. Since X is a complete metric space, there exists some $x^* \in X$ such that

$$\lim_{n \to \infty} x_n = x^*.$$

Furthermore, since ϕ is a contraction, it is continuous. So,

$$\begin{aligned} \phi(x^*) &= \phi\left(\lim_{n \to \infty} x_n \right) \\ &= \lim_{n \to \infty} \phi(x_n) \\ &= \lim_{n \to \infty} x_{n+1} \\ &= x^*, \end{aligned}$$

which shows that x^* is a fixed point of ϕ. Finally, we prove uniqueness: suppose that $x^*, y^* \in X$ are both fixed points. Then

$$\begin{aligned} d(x^*, y^*) &= d(\phi(x^*), \phi(y^*)) \\ &\leq c\, d(x^*, y^*) \\ \implies (1-c) d(x^*, y^*) &\leq 0 \end{aligned}$$

Since $c < 1$ and the metric is non-negative, this implies $d(x^*, y^*) = 0$, so $x^* = y^*$. To prove the last statement, note that since $\phi(x^*) = x^*$, repeated application of the contraction inequality gives

$$\begin{aligned} d(x_n, x^*) &= d(\phi^n(x_0), \phi^n(x^*)) \\ &\leq c^n d(x_0, x^*). \end{aligned}$$

□

Proposition A.6 (Completeness of the Value Function Space)**.** *Let $\mathcal{S}$ be a finite state space, and define the space of real-valued function on S to be*

$$\mathcal{V} = \{V : \mathcal{S} \to \mathbb{R}\}.$$

Consider $\mathcal{V}$ to be a metric space with supremum norm

$$\|V\|_\infty = \max_{s \in \mathcal{S}} |V(s)|,$$

is a complete metric space under the metric $d(V, W) = \|V - W\|_\infty$. ⌟

Proof. Let $(V_n)_{n \geq 0}$ be a Cauchy sequence in $\mathcal{V}$. By definiton of Cauchy sequences, then, for every $\varepsilon > 0$, there exists $N \in \mathbb{Z}_{>0}$ such that

$$\|V_n - V_m\|_\infty < \varepsilon$$

for all $n, m \geq N$. By definition of the supremum norm, then, for any $s \in \mathcal{S}$,

$$|V_n(s) - V_m(s)| \leq \|V_n - V_m\|_\infty < \varepsilon.$$

This means $(V_n(s))_{n\geq 0}$ is a Cauchy sequence in $\mathbb{R}$. Since $\mathbb{R}$ is complete, it converges to some $V(s) \in \mathbb{R}$. Note that we now have a function $V \in \mathcal{V}$. Taking $m \to \infty$ in the preceding inequality gives

$$|V_n(s) - V(s)| \leq \varepsilon$$

for every $s \in \mathcal{S}$ and $n \geq N$. So,

$$\|V_n - V\|_\infty \leq \varepsilon.$$

Therefore, $V_n \to V$ under the supremum norm, proving that $\mathcal{V}$ is complete. □