Lecture 2

Returns, Occupancy Measures, and the RL Objective

Goals of this lecture

1. Introduce the concept of discounted return and distinguish episodic vs. continuing tasks.
2. Define the RL objective as maximizing expected discounted return.
3. Introduce the discounted state-action occupancy measure and its use in RL.
4. Derive the marginal balance equation satisfied by the occupancy.
5. Show that stationary policies are sufficient by constructing one with equal performance.

2.1 Return and Task Types

We now formally introduce the concept of *return*, which quantifies the cumulative reward received by the agent and provides the basis for defining the reinforcement learning objective. We then describe two main categories of reinforcement learning tasks, emphasizing how the notion of return applies to each scenario.

Return. In reinforcement learning, the agent aims to maximize the cumulative reward it receives over time. Formally, we define the discounted cumulative reward, or *return*, from time step t onward as:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+1+k},$$

where $\gamma \in [0, 1]$ is the *discount factor*, controlling how much immediate rewards are favored over future rewards.

The structure and interpretation of this return depend significantly on the type of reinforcement learning task considered—specifically, whether the task has a natural ending or continues indefinitely.

Reinforcement learning tasks typically fall into two main categories: *episodic* and *continuing*. Both task types utilize the same definition of the return G_t, but they interpret it slightly differently based on whether interactions have natural termination points.

Episodic Tasks. An *episodic task* involves sequences of interactions called episodes. Each episode begins in an initial state (usually drawn from a given distribution) and ends when a particular terminal state or goal is reached. Examples of episodic tasks include games with well-defined endings or navigation tasks with specific target states.

In episodic tasks, the return from time step t until the termination of the episode at step T is naturally finite and commonly defined using a discount factor $\gamma \in [0, 1]$:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-t-1} R_T.$$

The termination time T is finite, but not necessarily fixed. Instead, represents a random variable describing the length of the episode (when the goal is reached or the game has ended).

Continuing Tasks. A *continuing task*, in contrast, does not have natural termination points, and interactions proceed indefinitely. Examples include ongoing process control or continuous monitoring scenarios.

Because the sequence of rewards continues indefinitely, continuing tasks must use a discount factor strictly less than 1 ($\gamma < 1$) to ensure the return remains finite and well-defined:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+1+k}.$$

Note that if rewards are bounded by some finite constant $\bar{r}$, i.e., $|R_t| \leq \bar{r}$ for all t, then having $\gamma < 1$ ensures the return G_t is finite, since it is bounded by:

$$|G_t| \leq \sum_{k=0}^{\infty} |\gamma^k R_{t+1+k}| \leq \bar{r} \sum_{k=0}^{\infty} \gamma^k = \frac{\bar{r}}{1-\gamma}.$$

Connecting Episodic and Continuing Tasks. Although episodic and continuing tasks differ in structure, the objective in both scenarios is identical: maximizing the expected discounted return G_t. In fact, episodic tasks can be viewed as special cases of continuing tasks by considering each terminal state as an absorbing state that transitions only to itself with reward zero. Under this interpretation, the discounted return in episodic tasks naturally matches the continuing-task formulation. This equivalence clarifies that a unified theoretical and algorithmic framework applies to both task types, as discussed in detail by Sutton and Barto [1, Section 3.4].

Other Task Settings. In addition to the discounted episodic and continuing tasks discussed above, there are two other common formulations worth mentioning.

The first is the *fixed-horizon* episodic setting, where each episode lasts exactly T steps, regardless of the agent's behavior. The return in this case is defined as a finite sum of rewards over the horizon:

$$G_t = R_{t+1} + R_{t+2} + \cdots + R_T = \sum_{k=0}^{T-t-1} R_{t+1+k}.$$

This setting is common in planning and control tasks with a fixed deadline or known duration.

The second is the *average-reward* formulation for continuing tasks, where the objective is to maximize the expected average reward per time step in the limit:

$$\bar{G} = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} R_{t+1}.$$

This objective is particularly useful in ongoing processes where discounting is either unnatural or undesirable, such as in operations research or steady-state control.

While both of these settings are important in specific domains, we will not focus on them in this course. Instead, we adopt the discounted return formulation, which unifies the treatment of both episodic and continuing problems and serves as the foundation for most modern reinforcement learning algorithms.

2.2 The Reinforcement Learning Objective

Having defined the concept of return, we can now formally state the primary goal of reinforcement learning.

Problem Formulation. Given a Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P)$ and an initial state distribution ρ, the agent's objective is to find a policy π that maximizes the expected discounted return. Formally, we seek:

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_\pi[G_0 \mid S_0 \sim \rho],$$

where Π denotes the set of all possible (history-dependent) policies, and the expectation $\mathbb{E}_\pi[\cdot]$ is taken over the distribution of trajectories induced by the policy π.

Occupancy Measure. To analyze and compare policies, it is useful to define a quantity that captures the frequency with which each state-action pair is visited under a given policy. For any policy $\pi \in \Pi$ and initial state distribution ρ, the *discounted state-action occupancy measure* is defined as:

$$\rho_\pi^\gamma(s, a) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \, \mathbb{P}_\pi(S_t = s, A_t = a \mid S_0 \sim \rho).$$

This defines a valid probability distribution over $\mathcal{S} \times \mathcal{A}$ whenever $\gamma < 1$, and it encodes the expected discounted visitation frequencies of state-action pairs under π. The marginal over states is given by:

$$\rho_\pi^\gamma(s) := \sum_{a \in \mathcal{A}} \rho_\pi^\gamma(s, a) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \, \mathbb{P}_\pi(S_t = s \mid S_0 \sim \rho),$$

which can be interpreted as the total discounted frequency of visits to state s.

The RL objective can then be written in terms of the occupancy measure as:

$$\mathbb{E}_\pi[G_0 \mid S_0 \sim \rho] = \frac{1}{1 - \gamma} \sum_{s,a} \rho_\pi^\gamma(s, a) \, r(s, a),$$

where $r(s, a) = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$ is the expected immediate reward.

Theorem 2.1 (Marginal Consistency of Discounted Occupancy)**.** *Let $\pi \in \Pi$ be any policy (possibly history-dependent), and let $\rho_\pi^\gamma(s, a)$ and $\rho_\pi^\gamma(s)$ denote its discounted occupancy measure and state marginal, respectively. Then $\rho_\pi^\gamma(s)$ satisfies the following equation:*

$$\rho_\pi^\gamma(s') = (1 - \gamma) \, \rho(s') + \gamma \sum_{s,a} \rho_\pi^\gamma(s, a) \, p(s' \mid s, a).$$

Proof. By definition, the discounted state marginal under π is:

$$\rho_\pi^\gamma(s') = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\,\mathbb{P}_\pi(S_t = s' \mid S_0 \sim \rho).$$

We split the sum into the $t=0$ term and the rest:

$$= (1-\gamma)\,\mathbb{P}_\pi(S_0 = s' \mid S_0 \sim \rho) + (1-\gamma)\sum_{t=1}^{\infty}\gamma^t\,\mathbb{P}_\pi(S_t = s' \mid S_0 \sim \rho).$$

Since $S_0 \sim \rho$, the first term equals $(1-\gamma)\,\rho(s')$. Reindexing the second sum:

$$= (1-\gamma)\,\rho(s') + \gamma(1-\gamma)\sum_{t=0}^{\infty}\gamma^t\,\mathbb{P}_\pi(S_{t+1} = s' \mid S_0 \sim \rho).$$

Now apply the law of total probability:

$$\mathbb{P}_\pi(S_{t+1} = s' \mid S_0 \sim \rho) = \sum_{s,a}\mathbb{P}_\pi(S_t = s, A_t = a \mid S_0 \sim \rho)\,p(s' \mid s,a).$$

Substitute this into the sum:

$$\gamma(1-\gamma)\sum_{t=0}^{\infty}\gamma^t\sum_{s,a}\mathbb{P}_\pi(S_t = s, A_t = a \mid S_0 \sim \rho)\,p(s' \mid s,a).$$

Rearrange sums:

$$= \gamma\sum_{s,a}\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t\mathbb{P}_\pi(S_t = s, A_t = a \mid S_0 \sim \rho)\right]p(s' \mid s,a).$$

Recognize the term in brackets as $\rho_\pi^\gamma(s',a')$, so we obtain:

$$\rho_\pi^\gamma(s') = (1-\gamma)\,\rho(s') + \gamma\sum_{s,a}\rho_\pi^\gamma(s,a)\,p(s' \mid s,a).$$

□

Corollary 2.1 (Fixed Point Equation for State Occupancy under Markov Policies)**.** *Let $\bar\pi \in \Pi_M$ be a stationary Markov policy and let ρ be the initial state distribution. Then the discounted state occupancy measure $\rho_{\bar\pi}^\gamma(s)$ satisfies:*

$$\rho_{\bar\pi}^\gamma(s') = (1-\gamma)\,\rho(s') + \gamma\sum_{s}\rho_{\bar\pi}^\gamma(s)p_{\bar\pi}(s'|s),$$

where $p_{\bar\pi}(s'|s) := \sum_a p(s' \mid s,a)\,\bar\pi(a \mid s)$.

Adequacy of Markov Policies. While the general problem formulation allows for arbitrary history-dependent policies, an important simplification arises when the environment dynamics satisfy the Markov property. In particular, one might question whether restricting attention to Markovian (stationary) policies $\pi \in \Pi_M$, which select actions based solely on the current state, results in any loss of optimality. Remarkably, under mild technical conditions, it can be rigorously established that the set of Markovian stationary policies Π_M is sufficient to achieve optimal performance:

$$\max_{\pi\in\Pi}\mathbb{E}_\pi[G_0 \mid S_0 \sim \rho] = \max_{\pi\in\Pi_M}\mathbb{E}_\pi[G_0 \mid S_0 \sim \rho].$$

This property, known as the *adequacy of Markov policies*, significantly simplifies both theoretical analysis and algorithmic design. Henceforth, our search for optimal policies will be confined to the class of Markovian stationary policies without loss of generality.

Theorem 2.2 (Adequacy of Markov Stationary Policies)**.** *Let $\pi \in \Pi$ be any (possibly history-dependent) policy, and let ρ be the initial state distribution. Define the* discounted state–action occupancy measure *of π as:*

$$\rho_\pi^\gamma(s,a) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\,\mathbb{P}_\pi(S_t = s, A_t = a \mid S_0 \sim \rho).$$

Let the state-marginal be $\rho_\pi^\gamma(s) := \sum_a \rho_\pi^\gamma(s,a)$. Now define a stationary Markov policy $\bar\pi$ by:

$$\bar\pi(a \mid s) := \begin{cases} \frac{\rho_\pi^\gamma(s,a)}{\rho_\pi^\gamma(s)} & \text{if } \rho_\pi^\gamma(s) > 0, \\ \text{any distribution on } \mathcal{A} & \text{otherwise.} \end{cases}$$

Then $\bar\pi \in \Pi_M$, and the following holds:

1. *The occupancy measures agree:* $\rho_{\bar\pi}^\gamma(s,a) = \rho_\pi^\gamma(s,a)$ *for all* (s,a).
2. *The expected returns are equal:* $\mathbb{E}_{\bar\pi}[G_0 \mid S_0 \sim \rho] = \mathbb{E}_\pi[G_0 \mid S_0 \sim \rho]$.

Proof. Let $\pi \in \Pi$ be any policy, and let $\rho_\pi^\gamma(s,a)$ be its discounted occupancy measure with state marginal $\rho_\pi^\gamma(s)$. Define $\bar\pi \in \Pi_M$ via:

$$\bar\pi(a \mid s) := \begin{cases} \frac{\rho_\pi^\gamma(s,a)}{\rho_\pi^\gamma(s)} & \text{if } \rho_\pi^\gamma(s) > 0, \\ \text{any distribution over } \mathcal{A} & \text{otherwise.} \end{cases}$$

Our goal is to show that $\rho_{\bar\pi}^\gamma(s,a) = \rho_\pi^\gamma(s,a)$, and therefore both policies yield the same expected return.

Step 1: Show that $\rho_\pi^\gamma(s)$ satisfies the fixed-point equation for $\bar\pi$. From Theorem 2.1, the marginal state occupancy under π satisfies:

$$\rho_\pi^\gamma(s') = (1-\gamma)\,\rho(s') + \gamma\sum_{s,a}\rho_\pi^\gamma(s,a)\,p(s' \mid s,a).$$

Now consider the fixed-point equation for the marginal occupancy under $\bar\pi$, as given in Corollary 2.1:

$$\rho_{\bar\pi}^\gamma(s') = (1-\gamma)\,\rho(s') + \gamma\sum_{s}\rho_{\bar\pi}^\gamma(s)\,p_{\bar\pi}(s' \mid s),$$

where $p_{\bar\pi}(s' \mid s) := \sum_a \bar\pi(a \mid s)\,p(s' \mid s,a)$.

We now check that $\rho_\pi^\gamma(s)$ satisfies this equation. Start with:

$$\begin{aligned}\sum_s \rho_\pi^\gamma(s)\,p_{\bar\pi}(s' \mid s) &= \sum_s \rho_\pi^\gamma(s)\sum_a \bar\pi(a \mid s)\,p(s' \mid s,a) \\ &= \sum_s\sum_a \rho_\pi^\gamma(s)\,\frac{\rho_\pi^\gamma(s,a)}{\rho_\pi^\gamma(s)}\,p(s' \mid s,a) \\ &= \sum_{s,a}\rho_\pi^\gamma(s,a)\,p(s' \mid s,a),\end{aligned}$$

where we used the definition of $\bar\pi(a \mid s)$. Substituting into the RHS of the fixed-point equation:

$$(1-\gamma)\,\rho(s') + \gamma\sum_{s,a}\rho_\pi^\gamma(s,a)\,p(s' \mid s,a),$$

which matches $\rho_\pi^\gamma(s')$ from Theorem 2.1. Therefore, $\rho_\pi^\gamma(s)$ satisfies the fixed-point equation for the Markov policy $\bar\pi$, and by uniqueness of the solution (under finite state/action spaces and $\gamma < 1$), we conclude:

$$\rho_{\bar\pi}^\gamma(s) = \rho_\pi^\gamma(s), \quad \forall s \in \mathcal{S}.$$

Step 2: Recover $\rho^\gamma_{\bar{\pi}}(s,a)$ from marginals. Using the fact that $\rho^\gamma_{\bar{\pi}}(s,a) = \rho^\gamma_{\bar{\pi}}(s)\,\bar{\pi}(a \mid s)$, we compute:

$$\rho^\gamma_{\bar{\pi}}(s,a) = \rho^\gamma_{\pi}(s)\,\frac{\rho^\gamma_{\pi}(s,a)}{\rho^\gamma_{\pi}(s)} = \rho^\gamma_{\pi}(s,a),$$

for all s, a such that $\rho^\gamma_{\pi}(s) > 0$, and by construction the same holds for $\rho^\gamma_{\pi}(s) = 0$.

Step 3: Conclude return equivalence. Since $\rho^\gamma_{\bar{\pi}}(s,a) = \rho^\gamma_{\pi}(s,a)$, the expected discounted return is the same:

$$\mathbb{E}_{\bar{\pi}}[G_0 \mid S_0 \sim \rho] = \frac{1}{1-\gamma}\sum_{s,a}\rho^\gamma_{\bar{\pi}}(s,a)\,r(s,a) = \frac{1}{1-\gamma}\sum_{s,a}\rho^\gamma_{\pi}(s,a)\,r(s,a) = \mathbb{E}_{\pi}[G_0 \mid S_0 \sim \rho].$$

□