Markov Decision Processes Lecture Notes 03

Objectives

Authors: Benjamin Van Roy *September 29, 2025*

1 Reward

For an MDP $(\mathcal{S}, \mathcal{A}, P)$, different policies can generate different dynamics. To choose among them, we need an objective that expresses preferences. We will specify objectives using reward functions.

A reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ maps state-action pairs to scalar rewards. This function assigns a numerical value $r(s, a)$ that indicates the desirability to being in a state s and taking action a. Let us provide a few examples.

1.1 Tetris

Recall the MDP $(\mathcal{S}, \mathcal{A}, P)$ specified in Lecture Notes 01 to model the game of Tetris. Each state s encodes a board configuration and the action set is $\mathcal{A} = \{\text{left}, \text{right}, \text{drop}, \text{none}\}$. There are multiple natural reward functions. One natural choice would be to take $r(s, a)$ to be the number of rows of the Tetris board that would be eliminated by taking the action a at state s. Another choice could be to assign a very large negative reward if (s, a) leads to termination of the game, and zero reward otherwise.

1.2 Queueing

In Lecture 01, we specified an MDP $(\mathcal{S}, \mathcal{A}, P)$ to model a queueing system. Each state is a queue length and the action indicates whether to apply a fast or slow mode of service. In such a system its natural to want to minimize wait times. One way to do that is by attributing a negative reward to having each customer wait at each time. The sum of these negative rewards at any given time is simply the queue length. Hence, we could take the reward $r(s, a)$ to be the number of waiting customers, which is the negative queue length $-s$.

1.3 Investment and Consumption

The investment process described in Lecture Notes 01 modeled the dynamics of wealth. While one could argue that an investor prefers more to less, a holistic model should view wealth as a means used to satisfy ends through its use for consumption. We now consider a model that endogenizes consumption decisions and attributes reward to consumption.

To keep things simple let us assume there is only one risky security. Hence, decisions allocate wealth between a riskless security, a risky security, and consumption. In particular, the MDP $(\mathcal{S}, \mathcal{A}, \mathcal{W}, f, \nu)$ has $\mathcal{S} = \mathbb{R}_+$, $\mathcal{A} = \{a \in \mathbb{R}^2 : a \geq 0, a_1 + a_2 \leq 1\}$, and $\mathcal{W} = \mathbb{R}_+$. The state expresses current wealth, the action provides fractions allocated to the risky asset and to current consumption. The disturbance is the return of the risky security. The return of the riskless security is a fixed constant z. The state update function is

$$f(s, a, w) = z(1 - a_2)s + (w - z)a_1 s.$$

One might consider attributing reward to the quantity of wealth consumed, in which case the reward function would be

$$r(s, a) = a_2 s.$$

However, as we will see in the next homework assignment, this incentivizes behavior that risks dire levels of future consumption. Taking reward instead to be

$$r(s, a) = u(a_2 s),$$

for a suitable concave increasing utility function u incentivizes more sensible behavior.

1.4 Revenue Management

Revenue management systems are deployed widely across the transportation and entertainment industry to manage automated sales, for example, of flight tickets. Here we consider a simple example.

Suppose a vendor sells tickets to a concert. They begin with S_0 tickets. While each ticket are indistinguishable to the vendor, they can increase their earnings via price differentiation. In particular, to segment customers, they sell two classes of tickets – standard and VIP – at prices $p_{\mathrm{std}} < p_{\mathrm{vip}}$. Customer's perceive a difference in status, perhaps signaled through red carpet access and glass of champagne offered to VIP ticket holders.

At each time t, the vendor receives a request for either a standard or VIP ticket, or neither. Let q_{std} and q_{vip} be the probabilities of the first two events and $1 - q_{\mathrm{std}} - q_{\mathrm{vip}}$, the third. The vendor's decision at each time indicates whether or not they are willing to accept a standard ticket request if one arrives. Any tickets not sold by time T become worthless since the concert takes place then and there will therefore be no further ticket requests.

This revenue management system can be modeled by an MDP $(\mathcal{S}, \mathcal{A}, P)$. The state and action spaces are $\mathcal{S} = \{0, 1, 2, \ldots\}^2$ and $\mathcal{A} = \{\text{accept}, \text{reject}\}$. At each time $t < T$, $S_{t,1}$ is the number of unsold tickets, $S_{t,2} = T - t$ is the remaining time, and A_t indicates whether a request for a standard ticket should be accepted. It is straightforward to provide expressions for transition probabilities. A natural reward function expresses the expected revenue over the next time step:

$$r(s,a) = \begin{cases} q_{\mathrm{std}} p_{\mathrm{std}} + q_{\mathrm{vip}} p_{\mathrm{vip}} & \text{if } s_2 > 0, a = \text{accept} \\ q_{\mathrm{vip}} p_{\mathrm{vip}} & \text{if } s_2 > 0, a = \text{reject} \\ 0 & \text{if } s_2 = 0. \end{cases}$$

1.5 LLM Math

Suppose we want to produce a language model that solves challenging mathematical problems. With such problems, it can be much easier to verify that a solution is correct based on the reasoning and final answer than it is to produce the solution. We formulate here an MDP $(\mathcal{S}, \mathcal{A}, P)$ that models dynamics and a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.

The action space is the set of tokens $\mathcal{A} = \{1, \ldots, M\}$. Each token could, for example, be an ASCII character, except for token M, which indicates the end of a message. The state space $\mathcal{S} = \cup_{t=0}^{\infty} \mathcal{A}^t$ is the set of token strings. The initial state S_0 is a string that describes a math problem and ends with token M. Each next state $S_{t+1} = f(S_t, A_t)$ is the concatenation of the previous string S_t with the next token A_t. A stationary policy π selects the next action according to $A_{t+1} \sim \pi(\cdot|S_t)$. The reward $r(S_t, A_t)$ is zero unless $A_t = M$ and the string S_t provides a math problem followed by a verifiably correct solution.

2 Return

An effective action should be chosen not only for its impact on the immediate reward $R_{t+1} = r(S_t, A_t)$ but also on subsequent rewards. To balance between rewards realized at different times, we need an operation that combines them to produce a single scalar metric that expresses desirability of the sequence of rewards. This scalar metric is often referred to as the *return*.

2.1 Total Return

With our revenue management system, for example, it is natural to simply sum rewards accrued through the time T. This gives rise to a metric

$$\sum_{t=0}^{T-1} R_{t+1}. \tag{1}$$

For the revenue management system, this is equivalent to the infinite sum $\sum_{t=0}^{\infty} R_{t+1}$ because $R_{t+1} = 0$ for $t \geq T$, since there are no ticket requests after time T.

An infinite sum of nonzero rewards is typically unbounded. In particular, $\sum_{t=0}^{\infty} R_{t+1}$ is typically $\pm\infty$ or undefined. While we may prefer sequences that lead to $+\infty$, there could be very many, and as such, this way of combining rewards does not sufficiently discriminate between trajectories: many that appear good at all will be assigned a score of $+\infty$.

2.2 Discounted Return

Discounting leads to an alternative approach for combining an infinite sequence of rewards. The discounted sum of rewards takes the form[1]

$$\sum_{t=0}^{\infty} \gamma^t R_{t+1}, \tag{2}$$

where $\gamma \in [0, 1)$ is a discount factor that expresses time preference. If rewards are bounded then the discounted sum is finite.

To offer one intepretation of this discounted sum, suppose the agent will have a finite but unknown lifespan T. In particular, let T be geometrically distributed, taking values in $\{1, 2, 3, \ldots\}$, with mean $1/(1-\gamma)$. Then, the average over possible lifespans is

$$\sum_{T=1}^{\infty} (1-\gamma)\gamma^{T-1} \sum_{t=0}^{T-1} R_{t+1} = \sum_{t=0}^{\infty} \gamma^t R_{t+1}. \tag{3}$$

Hence, discounting future rewards offers a natural way of weighting them based on the probability that the agent survives long enough to enjoy them.

Our model of investment and consumption offers another context where discounting may be natural. One may discount to model a random lifespan after which wealth cannot be consumed. Or one may discount to reflect an individual's personal preferences. For example, an individual could enjoy consuming more sooner rather than later.

2.3 Average Return

MDPs are often used to model processes with very small time steps and where rewards remain relevant over a long duration. For example, a typical Internet switch may operate at around 100MHz, with up to a hundred million actions executed each second. And the throughput and delay achieved matter throughout a very long duration. While such a system and its performance objectives can be modeled in terms of an MDP and a reward function, how should rewards be combined over time? If we use the discounted return, the discount factor ought to be very close to one. since it is not clear what specific value to choose, we might as well consider the limit as the discount factor approaches one. This motivates the average return:[2]

$$\lim_{\gamma \uparrow 1} (1-\gamma) \sum_{t=0}^{\infty} \gamma^t R_{t+1} = \lim_{t \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} R_{t+1}. \tag{4}$$

[1] This limit is guaranteed to exist for a countable-state MDP and bounded reward function. Beyond that, other technical conditions are required.

[2] These limits are guaranteed to exist and be equal for a finite-state MDP under a stationary policy. Beyond that, other technical conditions are required.

The average return offers a useful objective in many environments. For example, for our queueing model, minimizing the average queue length seems reasonable. And for our inventory model, minimizing average cost seems reasonable.

But there are contexts where average return seems like a natural choice but does not quite suffice. To offer an example, let us revisit Tetris. Average return offers a useful metric if the game lasts indefinitely. But if it does not then the average return will be zero. And, among trajectories that earn zero average return, we would naturally prefer those that accumulate greater reward. The average return does not distinguish between such trajectories.

To remedy this, it is common to use a second metric, which we will call the *relative return*. The relative return is defined by

$$\lim_{\gamma \uparrow 1} \sum_{t=0}^{\infty} \gamma^t (R_{t+1} - \lambda), \tag{5}$$

where λ is the average return. Intuitively, the relative return measures the extent to which rewards are above or below the long-term average before the running average converges. The relative reward serves to break ties. When two trajectories share the same average return, the one with greater relative reward is preferred.

If the Tetris game ends then $\lambda = 0$ and the relative return is the total return. If the game never ends then it is natural to prefer trajectories with larger average return and, if required, break ties based on relative return.

3 Expected Return

We have discussed forms of return, which combine rewards across a trajectory. While returns offer metrics for choosing between trajectories they do not suffice when it comes to choosing between policies. This is because rewards are realized *after* a policy is chosen. The choice must be based on expectations about the future.

3.1 Expected Total Return

We could, for example, take the expectation of total return:

$$\mathbb{E}_{S_0,\pi}\left[\sum_{t=0}^{T-1} R_{t+1}\right]. \tag{6}$$

This represents an integral over probability-weighted trajectories. The subscripts S_0 and π specify the initial state and the policy used to select actions. This expectation depends on the initial state and policy. The expected total return *can* be used as an objective for selecting between policies. Each policy yields a different expectation, and larger values are preferred.

3.2 Expected Discounted Return

Taking the expectation of discounted return leads to a second objective:

$$\mathbb{E}_{S_0,\pi}\left[\sum_{t=0}^{\infty} \alpha^t R_{t+1}\right]. \tag{7}$$

3.3 Gain and Bias

Finally, another useful objective is the expected average return, also called the *gain*:

$$\lim_{T \to \infty} \mathbb{E}_{S_0,\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} R_{t+1}\right]. \tag{8}$$

A policy that maximizes the gain is said to be *gain-optimal.* The expected relative return, also called the *bias*, is defined by

$$\lim_{\gamma \uparrow 1} \mathbb{E}_{S_0,\pi}\left[\sum_{t=0}^{\infty} \gamma^t (R_{t+1} - \lambda_{S_0,\pi})\right], \tag{9}$$

where $\lambda_{S_0,\pi}$ is the gain. We can define a lexicographic objective that is optimized by choosing from among gain-optimal policies one that maximizes the bias. Such a policy is referred to as *bias-optimal.* The bias serves to break ties between gain-optimal policies.

3.4 Concrete Notation

We have expressed expected returns via somewhat abstract notation. Each expectation can be approximated via simulating trajectories and averaging returns. When state and action spaces are finite and policies are stationary, such expectations can alternatively be computed via matrix operations. Expressing the expectations in terms of more concrete matrix notation makes that clear. We now define such notation.

Consider an MDP $(\mathcal{S}, \mathcal{A}, P)$ with finite $\mathcal{S}$ and $\mathcal{A}$. Without loss of generality, let $\mathcal{S} = \{1, \dots, |\mathcal{S}|\}$ and $\mathcal{A} = \{1, \dots, |\mathcal{A}|\}$. For any stationary policy π, let P_π be a matrix with elements $P_{\pi s s'} = \sum_{a \in \mathcal{A}} \pi(a|s) P_{ass'}$. Let r_π be a vector with components $r_{\pi s} = \sum_{a \in \mathcal{A}} \pi(a|s) r(s,a)$. We then have

$$\mathbb{E}_{S_0,\pi}\left[\sum_{t=0}^{T-1} R_{t+1}\right] = \sum_{t=0}^{T-1} \left(P_\pi^t r_\pi\right)_{S_0}, \tag{10}$$

$$\mathbb{E}_{S_0,\pi}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1}\right] = \sum_{t=0}^{\infty} \gamma^t \left(P_\pi^t r_\pi\right)_{S_0}, \tag{11}$$

$$\lim_{T \to \infty} \mathbb{E}_{S_0,\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} R_{t+1}\right] = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \left(P_\pi^t r_\pi\right)_{S_0}, \tag{12}$$

$$\lim_{\gamma \uparrow 1} \mathbb{E}_{S_0,\pi}\left[\sum_{t=0}^{\infty} \gamma^t (R_{t+1} - \lambda_{S_0,\pi})\right] = \lim_{\gamma \uparrow 1} \sum_{t=0}^{\infty} \gamma^t \left(\left(P_\pi^t r_\pi\right)_{S_0} - \lambda_{S_0,\pi}\right), \tag{13}$$

where $\lambda_{S_0,\pi}$ is the gain.

4 State-Action Probabilities

Expressing expected returns in terms of state-action probabilities can offer additional insight. In this section, we will study such expressions for infinite-horizon objectives. We restrict attention to MDPs with finite state and action spaces and policies that are stationary. Concepts we study extend to infinite state and action spaces, under suitable technical assumptions. The restriction to stationary policies is warranted because, as we will see in future lectures, in most contexts of interest, for each of our infinite-horizon objectives, there exists an optimal policy that is stationary.

4.1 Average Return

In this section, we will make a simplifying assumption that each stationary policy π induces a Markov process $(\mathcal{S}, P_\pi)$ with a single recurrent class of states. While ideas extend more generally, they become more complicated in the absence of this assumption.

4.1.1 A Balance Equation for State-Action Probabilities

Under our assumption, the gain $\lambda_{S_0,\pi}$ does not depend on S_0, so we write it simply as λ_π. Further, each stationary policy π induces a steady-state distribution μ_π that uniquely solves

$$\mu(s') = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \mu(s) P_{\pi s s'} \tag{14}$$

within $\Delta_{\mathcal{S}}$. The gain under π can then be written as

$$\lambda_\pi = \sum_{s\in\mathcal{S}} \mu_\pi(s) \sum_{a\in\mathcal{A}} \pi(a|s) r(s,a) = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \mu_\pi(s,a) r(s,a), \tag{15}$$

where, with some abuse of notation, we define *state-action probabilities* $\mu_\pi(s,a) = \mu_\pi(s)\pi(a|s)$.

The state-action probabilities μ_π uniquely solve a balance equation

$$\sum_{a'\in\mathcal{A}} \mu(s',a') = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \mu(s,a) P_{ass'} \qquad \forall s' \in \mathcal{S} \tag{16}$$

within $\Delta_{\mathcal{S}\times\mathcal{A}}$. Further, any μ that solves this equation is induced by some policy π. Hence, there is a policy π with gain λ_π if and only if there is a solution to (16) such that $\lambda_\pi = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \mu(s,a) r(s,a)$.

4.1.2 Geometry of State-Action Probabilities

As we have explained, μ solves the balance (16) if and only if there is a stationary policy π such that $\mu = \mu_\pi$. The set of solutions is the intersection of a unit simplex and solutions to a linear system of equations. Hence, it is defined by linear constraints and thus a polytope. Let us denote this polytope by $\mathcal{P}$.

It can be shown that μ is a vertex of $\mathcal{P}$ if and only if $\mu = \mu_\pi$ for some deterministic stationary policy π. Hence, vertices map to deterministic policies while other elements of the polytope map only to stochastic policies.

Because $\mu \in \mathcal{P}$ if and only if $\mu = \mu_\pi$ for some stationary policy π, we have

$$\max_\pi \lambda_\pi = \max_{\mu\in\mathcal{P}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \mu(s,a) r(s,a). \tag{17}$$

Hence, the maximal gain can be obtained by solving a linear program. Further, if we were to use the simplex method, that would deliver a vertex solution μ_*. This vertex is equal to μ_π for some deterministic stationary policy π. Hence, under our assumptions there must always be a deterministic stationary policy that is optimal.

4.2 Discounted Return

Concepts we have presented in the context of average return extend to discounted return. In particular, maximizing discounted return is equivalent to maximizing average return with a modified transition matrix: $\gamma P + (1-\gamma)\mathbf{1}\mathbf{1}_{S_0}^\top$. This gives rise to a modified balance equation:

$$\sum_{a'\in\mathcal{A}} \mu(s',a') = \gamma \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \mu(s,a) P_{ass'} + (1-\gamma)\mathbf{1}_{S_0}(s') \qquad \forall s' \in \mathcal{S}. \tag{18}$$

For any stationary policy π, there exists a unique solution to this balance equation. This solution is given by

$$\mu_\pi(s,a) = (1-\gamma)\mathbb{E}_{S_0,\pi}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{1}_{s,a}(S_t, A_t)\right]. \tag{19}$$

The expected discounted return of a policy π can be written as

$$(1-\gamma)\mathbb{E}_{S_0,\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(S_t,A_t)\right] = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\mu_\pi(s,a)r(s,a). \tag{20}$$

The maximal expected discounted return can be obtained by solving a linear program:

$$\max_{\mu\in\mathcal{P}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\mu(s,a)r(s,a), \tag{21}$$

where the polytope $\mathcal{P}$ is comprised of solutions to balance equation (18) that lie within the $\Delta_{\mathcal{S}\times\mathcal{A}}$. An element $\mu\in\mathcal{P}$ is a vertex if and only $\mu=\mu_\pi$ for some deterministic policy π.

5 Pathological Cases

What we have covered in this lecture suggests that, given an MDP and a reward function, there is always a stationary deterministic policy that is optimal. While that is *typically* the case, it is not *always* the case. For example, there are pathological cases where no policy maximizes expected return. These pathological cases are of interest to mathematicians who aim to develop coherent foundations for the field. But they almost never arise in practical contexts. To give a sense for such pathological cases, let us discuss a couple in this section.

A first example illustrates how there may be no policy that maximizes expected return. While we could construct such an example with other notions of return as well, we will focus on total return. Consider an MDP $(\mathcal{S},\mathcal{A},P)$ with a single state $\mathcal{S}=\{1\}$ and a countably infinite action space $\mathcal{A}=\{1,2,\ldots\}$. Let the reward function be defined by $r(s,a)=1-1/a$. Consider an objective of maximizing expected total return over a single time period: $\mathbb{E}_{S_0,\pi}[r(S_0,A_0)]$. For any action a, the action $a'=a+1$ earns larger reward. Similarly, forany policy π, the policy π' defined by $\pi'(1|h)=0$ and $\pi'(a+1|h)=\pi(a|h)$, for $a\in\mathcal{A}$, earns larger expected reward. It follows that no policy maximizes expected return.

In a second example, involving the same MDP and reward function, there is an optimal policy but no stationary policy that is optimal. An optimal policy is given by $\pi(a|h)=\text{length}(h)+1$. In other words, select actions $A_0=1$, $A_1=2$, $A_2=3$, and so on. This policy attains a gain of $\lambda_{S_0,\pi}=1$, which must be optimal because that is the largest possible per-time-step reward. However, it is not a stationary policy. For any stationary policy π, there is another stationary policy π' defined by $\pi'(1|s)=0$ and $\pi'(a+1|s)=\pi(a|s)$ that attains larger gain.