IEOR 8100: Reinforcement learning 5-6 sessions

Lecture 1: Introduction: Reinforcement Learning and MDPs

By Shipra Agrawal

1 Introduction to reinforcement learning

Reinforcement learning is characterized by an agent continuously interacting and learning from a stochastic environment. It is essentially the science of making sequential decisions under uncertainty, by learning from the response to the past decisions.

Imagine a robot is moving around in the physical world, and wants to go from point A to B. How should the robot move its limbs so that it can eventually learn to walk and reach point B quickly? More generally, "how" should an automated agent interact with the environment, what actions should it take now, so that it is able to learn more about the environment and is eventually successful in its goals.

To do so, the robot can try different ways of moving its legs, learns from its successful motion as well as from its falls, and finally find the most effective way to walk and reach the target. Reinforcement learning is a branch of artificial intelligence that formalizes this trial-and-error method of learning.

Reinforcement learning sits at the intersection of many disciplines of science, namely:

- Optimal control (Engineering)
- Dynamic Programming (Operations Research)
- Reward systems (Neuro-science)
- Classical/Operant Conditioning (Psychology)

In all these different fields of study, there is a branch that is trying to study the same problem as reinforcement learning – essentially the problem of how to make optimal sequential decisions. In engineering it is the problem of finding optimal control, in Operations research it is studied under Dynamic programming. The algorithmic principles behind reinforcement learning are motivated from the natural phenomena behind human decision making – in simple words that "rewards provide a positive reinforcement for an action": this phenomena is studied in psychology as conditioning and in neuroscience as reward systems.

Key characteristics

Lets look at a few key features that make reinforcement learning different from other paradigms of optimization, and other machine learning methods like supervised learning. These characteristics also make RL a powerful model of learning for a variety of application domains.

- **Lack of a supervisor:** One of the main characteristic of RL is that there is no supervisor – no labels telling us the best action to take, only rewards as signals to enforce some actions more than others. For example, for a robot trying to walk, there is no supervisor telling the robot if the actions it took were a good way to walk and reach its target. But, it does get some signals in the form of immediate effect of its actions - moving forward or falling down – which it can use to guide its behavior.
- **Decisions affect data:** A further distinction between labels in supervised learning and rewards in RL is that the same action can have different rewards depending on the "state" of the environment and the agent, which itself depends on the past actions. For example, moving a leg in a certain way will generate different effects depending on the stance and position of the robot. And, this stance or position depends on the movements of the robot in the past. Therefore, the data points that the agent collects during the course of learning are not

independent of each other. In particular, they are a function of the agents' own past actions, which the agent may decide based on its past observations. This is very unlike other ML paradigms like supervised learning, where the training examples are often assumed to be independent of each other, or at least oblivious to the learning agent's actions. In RL, essentially the "quality of training data" available at any point depends on the agent's actions while the data was collected - if the robot explored a lot of different limb movements even if at the cost of falling down a lot, the collected data (observations) will be richer for learning. This highlights a potential tradeoff betweem immediate rewards and information, referred to as the exploration vs. exploitation tradeoff.

- **Delayed feedback:** The feedback is often delayed in RL settings: the effect of an action may not be entirely visible instantaneously, as the action may severely affect the reward signal many steps later. In the robot example, an aggressive movement of legs may look good right now as it may seem to make the robot go quickly towards the target, but a sequence of limb movements later you may realize that that aggressive movement made the robot fall. Such delayed or long-term effects makes it difficult to attribute credit and reinforce a good move whose effect may be seen only many steps and many moves later. This challenge is also referred to as the "credit assignment problem".

Examples

Lets concretize this discussion with some examples. We already discussed the example of a robot trying to walk. Here are a few others:

- **Automated vehicle control:** Imagine trying to fly an unmanned helicopter, learning to perform stunts with it. Again, no one will tell you if a particular way of moving the helicopter was good or not, you may get signals in form of how the helicopter is looking in the air, that it is not crashing, and in the end you might get a reward – for example a high reward for a good stunt, negative reward for crashing. Using these signals and reward, a reinforcement learning agent would learn a sequence of maneuvers just by trial and error.

- **Learning to play games:** Some of the most famous successes of reinforcement learning have been in playing games. You might have heard about Gerald Tesauro's reinforcement learning agent defeating world Backgammon Champion, or Deepmind's Alpha Go defeating the world's best Go player Lee Sedol, using reinforcement learning. A team at Google Deepmind built an RL system that can learn to play suite of Atari games from scratch by just by playing the game again and again, trying out different strategies and learning from their own mistakes and successes. This RL agent uses just the pixels of game screen as state and score increase as reward, and is not even aware of the rules of the game to begin with!

- **Medical treatment planning:** A slightly different but important application is in medical treatment planning. Here, the problem is to learn a sequence of treatments for a patient based on the reactions to the past treatments, and current state of the patient. Again, while you observe reward signals in the form of the immediate effect of a treatment on patient's condition, the final reward is whether the patient could be cured or not, and that can be only observed later (after several treatments/months/years). The trials are very expensive and slow in this case, and need to be carefully performed to achieve most efficient learning possible.

- **Chatbots:** Another popular application is chatbots: you might have heard of Microsoft's chatbots Tay and Zo, or intelligent personal assistants like Siri, Google Now, Cortana, and Alexa. All these agents try to make a conversation with a human user. What are conversations? They are essentially a sequence of sentences exchanged between two people. Again, a bot trying to make a conversation may occasionally receive encouraging signals if it is making a good conversation, or negative signals in the form of human user leaving the conversation or getting annoyed. A reinforcement learning agent attempts to use this feedback to learn how to make good conversations by trial and error: by reinforcing the patterns of conversations with good feedback, and steering away from the patterns of those where the user was not happy. And after many many conversations, you may have a chatbot which has learned the right thing to say at the right moment!

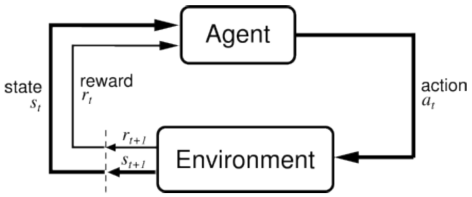


Figure 1: Figure taken from Sutton and Barto 1998

2 Introduction to MDP: the optimization/decision model behind RL

Markov decision processes or MDPs are the stochastic decision making model underlying the reinforcement learning problem. Reinforcement learning is essentially the problem when this underlying model is either unknown or too difficult (large) to solve in order to find an optimal strategy in advance. Therefore, instead the reinforcement learning agent learns and optimizes the model through execution and simulation, continuously using feedback from the past decisions to learn the underlying model and reinforce good strategies. But, more on that later, first lets understand what is a Markov decision process.

In this sequential decision making model, the agent needs to make decisions in discrete rounds $t = 1, 2, \ldots,$. In every round, all the relevant information from the past, is captured in the state of the process. The definition of 'state' depends on the problem, and is part of the modeling process. Let's consider again the example of a robot who is trying to move from point A to B. The current state of the robot in this example could be a combination of the location of the robot, the stance of the robot: whether it is standing, sitting, or moving, and its current velocity if it is moving. The decision maker, which in our example the robotic controller, observes the current state and takes an action, for example whether to lift a leg, or to move a limb of the robot forward. The agent then observes the reward signal and the transition to the next state, which depend both on the action taken and the state in which it was taken. For example, aggressively moving a leg may be a good action when the robot is walking or running, it will produce a positive reward signal and next state will also be a desirable state – the robot would have moved closer to the target. However, when the robot is still say in a standing position, the same aggressive action may make the robot fall down.

The defining property of MDPs is the Markov property which says that the future is independent of the past given the current state. This essentially means that the state in this model captures all the information from the past that is relevant in determining the future states and rewards.

Formal definition. A Markov Decision Process (MDP) is specified by a tuple (S, s_1, A, P, R, H), where S is the set of states, s_1 is the starting state, A is the set of actions. The process proceeds in discrete rounds $t = 1, 2, \ldots, H$, starting in the initial state s_1. In every round, t the agent observes the current state $s_t \in S$, takes an action $a_t \in A$, and observes a feedback in form of a reward signal $r_{t+1} \in \mathbb{R}$. The agent then observes transition to the next state $s_{t+1} \in S$.

The Markov property mentioned earlier is formally stated as the following two properties: firstly, the probability of transitioning to a particular state depends only on current state and action, and not on any other aspect of the history. The matrix $P \in [0, 1]^{S \times A \times S}$ specifies these probabilities. That is,

$$\Pr(s_{t+1} = s' | \text{ history till time } t) = \Pr(s_{t+1} = s' | s_t = s, a_t = a) = P(s, a, s')$$

And secondly, the reward distribution depends only on the current state and action. So, that the expected reward at time t is a function of current state and action. A matrix R specifies these rewards.

$$\mathbb{E}[r_{t+1} | \text{ history till time } t] = \mathbb{E}[r_{t+1} | s_t = s, a_t = a] = R(s, a)$$

In some problems, a different reward r_{t+1} may be specified for every triple s_t, a_t, s_{t+1}. This is equivalent to the above model. Let $R(s, a, s')$ be the expected (or deterministic) reward when action a is taken in state s and

transition to state s' is observed. Then, we can obtain the same model as above by defining

$$R(s,a) = \mathbb{E}[r_{t+1}|s_t = s, a_t = a] = \mathbb{E}_{s' \sim P(s,a)}[R(s,a,s')]$$

Policy A policy specifies what action to take at any time step. A history dependent policy at time t is a mapping from history till time t to an action. A Markovian policy is a mapping from state space to action action $\pi : S \to A$. Due to Markovian property of the MDP, it suffices to consider Markovian policies. In particular, for any history dependent policy π, there exists a Markovian (randomized) policy π' such that $\Pr(s_t = s, a_t = a|s_1)$ is same for both the policies. This is formally proven in Theorem 5.5.1 of Puterman [1994]. (The proof is constructive, see Example 5.5.1 in Puterman [1994] for an illustration of the construction of such a Markovian policy from a history dependent policy.) Given this observation, from hereon in this text, a policy refers to a Markovian policy.

A deterministic policy $\pi : S \to A$ is mapping from any given state to an action. A randomized policy $\pi : S \to \Delta^A$ is a mapping from any given state to a distribution over actions. Following a policy π_t at time t means that if the current state $s_t = s$, the agent takes action $a_t = \pi_t(s)$ (or $a_t \sim \pi_t(s)$ for randomized policy). In general (with some abuse of terminology), a non-stationary policy refers to a sequence of policies $(\pi_1, \pi_2, \ldots, \pi_t, \ldots)$. A stationary policy then refers to a static sequence $(\pi, \pi, \ldots, \pi)$, i.e., $\pi_t = \pi$ for all rounds $t = 1, 2, \ldots$.

Any stationary policy π defines a Markov chain, or rather a Markov reward process (MRP), that is, a Markov chain with reward associated with every transition. The transition probability vector and reward for this MRP in state s is given by $\Pr(s'|s) = P^\pi_s, \mathbb{E}[r_t|s] = r^\pi_s$, where P^π is an $S \times S$ matrix, and r^π is an S-dimensional vector defined as:

$$P^\pi_{s,s'} = \mathbb{E}_{a \sim \pi(s)}[P(s,a,s')], \forall s, s' \in S$$
$$r^\pi_s = \mathbb{E}_{a \sim \pi(s)}[R(s,a)]$$

The stationary distribution (if exists) of this Markov chain when starting from state s_1 is also referred to as the stationary distribution of the policy π, denoted by d^π:

$$d^\pi(s) = \lim_{t \to \infty} \Pr(s_t = s|s_1, \pi)$$

Goals. The tradeoffs between immediate reward vs. future rewards of the sequential decisions and the need for planning ahead is captured by the goal of the Markov Decision Process. At a high level, the goal is to maximize some form of cumulative reward. Some popular forms are total reward, average reward, or discounted sum of rewards.

- **Finite horizon MDP**: Here, actions are taken for $t = 1, \ldots, H$ where H is a finite horizon. The total (discounted) reward criterion is simply to maximize the expected total (discounted) rewards in an episode of length H. (In reinforcement learning context, when this goal is used, the MDP is often referred to as an episodic MDP.) For discount $0 \leq \gamma \leq 1$, the goal is to maximize

$$\mathbb{E}[\sum_{t=1}^{H} \gamma^{t-1} r_t|s_1]$$

- **Infinite horizon MDP:**
  - Expected total discounted reward criteria: The most popular form of cumulative reward is expected discounted sum of rewards. This is an asymptotic weighted sum of rewards, where with time the weights decrease by a factor of $\gamma < 1$. This essentially means that the immediate returns more valuable than those far in the future.

$$\lim_{T \to \infty} \mathbb{E}[\sum_{t=1}^{T} \gamma^{t-1} r_t|s_1]$$

– Expected total reward criteria: Here, the goal is to maximize

$$\lim_{T\to\infty} \mathbb{E}[\sum_{t=1}^{T} r_t|s_1]$$

The limit may not always exist or be bounded. We are only interested in cases where above exists and is finite. This requires restrictions on reward and/or transition models. Interesting cases include the case where there is an undesirable state, the reward after reaching that state is 0. For example, end of of a computer game. The goal would be to maximize the time to reach this state. (A minimization version of this model is where there is a cost associated with each state and the game is to minimize the time to reach winning state, called the shortest path problem).

– Expected average reward criteria: Maximize

$$\lim_{T\to\infty} \mathbb{E}[\frac{1}{T}\sum_{t=1}^{T} r_t|s_1]$$

Intuitively, the performance in a few initial rounds does not matter here, what we are looking for is a good asymptotic performance. This limit may not always exist. Assuming bounded rewards and finite state spaces, it exists under some further conditions on policy used.

Discounted sum of rewards is one of the most popular forms of goal in MDP for many reasons: it is mathematically convenient as it is always finite and avoids the complications due to infinite returns. Practically, depending on the application, immediate rewards may indeed be more valuable. Further, often uncertainty about far future are not well understood, so you may not want to give as much weight to what you think you might earn far ahead in future. The discounted reward criteria can also be seen as a soft version of finite horizon, as the contribution of reward many time steps later is very small. As you will see later, discounted reward MDP has many desirable properties for iterative algorithm design and learning. Due to these reasons, often the practical approaches which actually execute the MDP for finite horizon, use policies, algorithms and insights from infinite horizon discounted reward setting.
Remark on existence of limits. Assuming bounded rewards, finite state space and action space, the above limit always exists for the discounted reward case. For the average case, the limit exists for all stationary policies [see Puterman [1994] Proposition 8.1.1]. See Example 8.1.1 for an example of MDP and a non-stationary policy for which the limit does not exist in average reward case. For further discussion on existence of these limits in expected total reward case, see Chapter 5.1 and 5.2 of Puterman [1994].

Gain of the MDP. Gain (roughly the 'expected value objective' or formal goal) of an MDP when starting in state s_1 is defined as (when supremum exists):

- episodic MDP:

$$\rho(s_1) = \sup_{\{\pi_t\}} \mathbb{E}[\sum_{t=1}^{H} \gamma^{t-1} r_t|s_1]$$

- Infinite horizon expected total reward.

$$\rho(s_1) = \sup_{\{\pi_t\}} \lim_{T\to\infty} \mathbb{E}[\sum_{t=1}^{T} r_t|s_1]$$

- Infinite horizon discounted sum of rewards.

$$\rho(s_1) = \sup_{\{\pi_t\}} \lim_{T\to\infty} \mathbb{E}[\sum_{t=1}^{T} \gamma^{t-1} r_t|s_1]$$

- infinite horizon average reward:

$$\rho(s_1) = \sup_{\{\pi_t\}} \lim_{T\to\infty} \mathbb{E}[\frac{1}{T}\sum_{t=1}^{T} r_t|s_1]$$

Here, expectation is taken with respect to state transition and reward distribution, supremum is taken over all possible sequence of policies for the given MDP. It is also useful to define gain ρ^π of a stationary policy π, which is the expected (total/total discounted/average) reward when policy π is used in all time steps. For example, for infinite horizon average reward:

$$\rho^\pi(s_1) = \lim_{T\to\infty} \mathbb{E}[\frac{1}{T}\sum_{t=1}^{T} r_t|s_1]$$

where $a_t = \pi(s_t), t = 1, \ldots, T$.

Optimal policy. Optimal policy is defined as the one that maximizes the gain of the MDP. Due to the structure of MDP it is not difficult to show that it is sufficient to consider Markovian policies. Henceforth, we consider only Markovian policies.

For infinite horizon MDP with average/discounted reward criteria, a further observation that comes in handy is that such a MDP always has a stationary optimal policy, whenever optimal policy exists. That is, there always exists a fixed policy so that taking actions specified by that policy at all time steps maximizes average/discounted/total reward. The agent does not need to change policies with time. This insight reduces the question of finding the best sequential decision making strategy to the question of finding the best stationary policy.

The results below assume finite and countable states and actions, and bounded rewards.

Theorem 1 (Puterman [1994], Theorem 6.2.7)**.** *For any infinite horizon discounted MDP, there always exists a deterministic stationary policy π that is optimal.*

Theorem 2 (Puterman [1994], Theorem 7.1.9)**.** *For any infinite horizon expected total reward MDP, assuming for all policies either the upper or lower limit of the total reward objective exists, then there exists a deterministic stationary policy π that is optimal.*

Theorem 3 (Puterman [1994], Theorem 8.1.2)**.** *For infinite horizon average reward MDP, there always exist a stationary (possibly randomized) policy which is an optimal policy .*

Therefore, for infinite horizon MDPs, optimal gain:

$$\rho^*(s) = \max_{\pi:\text{Markovian stationary}} \rho^\pi(s)$$

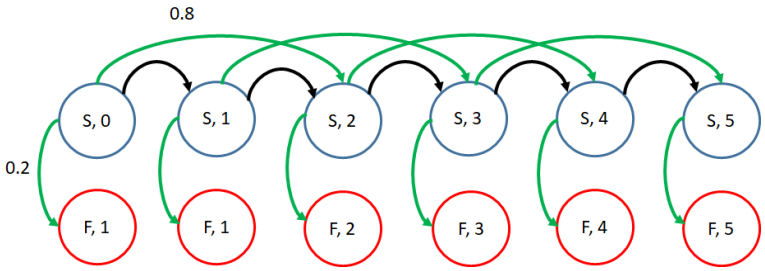
These results imply that the optimal solution space is simpler for infinite horizon case, and make infinite horizon an attractive model even when the actual problem is finite horizon but the horizon is long. Even when such a result on optimality of stationary policy is not available, 'finding the best stationary policy' is often used as an alternate convenient and more tractable objective, instead of finding the optimal policy which may not exist or may not be stationary in general.

Solving an MDP, finding optimal policy. Solving or optimizing an MDP means finding a strategy for the agent to choose actions in such a way so as to maximize the stated form of cumulative reward. Note that an action not only determines the current reward, but also future states and therefore future rewards. So, the agent needs to choose these actions or policy strategically in order to optimize the overall cumulative reward. The rest of the lecture will develop formal constructs necessary to design algorithmic solutions for solving this optimization problem. But, let's first look at some examples.

3 Examples

Example 1. Lets formulate a simple MDP for a robot moving on a line. Let's say there are only three actions available to this robot: walk or run or stay. Walking involves a slow limb movement, which allows the robot to move one step without falling. Running involves an aggressive limb movement which may allow the robot to move two steps forward, but there is a 20% chance to fall. Once the robot falls, it cannot get up. The goal is to move forward quickly and as much as possible without falling.

We can model this as MDP. We define state of the robot as a combination of its Stance: whether it is standing upright or has fallen down, denoted here as S or F, and its location on the line, represented here as $0, 1, 2, 3, 4, 5, \ldots$. So, this state $(S, 1)$ for example means that the robot is upright at location 1, where as this state (F,2) means that the robot has fallen down at location 2. The robot starts in a standing state at the beginning of the line, that is at state (S,0). Action space consists of three actions: walk , run, and stay. The state transition depends on current state and action. Walking in a standing state always transfers the robot to a standing state at a location one step ahead(this transition on taking walk action is represented here by these black arrows). So, by walking the robot can always move up by one step. On the other hand, by taking the second action of running or an aggressive limb movement, a robot in a standing state may move by 2 steps at a time (shown here by these green arrows), but there is also a 20% chance of falling and transitioning to a Fallen state. In fallen state, there is no effect of any action, the robot is stuck. Stay action keeps the robot in the current state.



As is often the case in applications of MDPs or more generally, reinforcement learning, the rewards and goals are not exogeneously given but are also an important part of the application modeling process. Different settings will lead to different interpretations of the problem. Lets say the reward is the number of steps the agent moves as a result of an action in the current state. Now, If the goal is set to be the total reward (infinite horizon), then the agent should just walk, because the aim is to move as many steps as possible, so moving quickly is not important, and the robot should not take the risk of falling by running. The total reward is infinite. But if the goal is set as discounted reward, then it is also important to move more steps initially and gather more reward quickly before the discount term becomes very small, so it may be useful to run (depending on how small the discount factor γ is). One can also set reward to be 0 for all the states except the final destination, say (S,5), where the reward is 1. In that case, the discounted sum of rewards would be simply γ^τ if (S,5) is reached at time τ, so the agent will want to minimize τ, the time to reach the end without falling, and therefore may want to move aggressively at times.

Another important point to note from this example, is that in an MDP an action has long term consequences. For instance, it may seem locally beneficial to run on state (S,0) here because, even with some chance of falling, the 2 steps gain at 80% chance means that expected immediate reward is .8*2 = 1.6, which is more than the expected immediate reward of 1 step that can be earned by walking, but that greedy approach ignores all the reward you can make in future if you don't fall. Finding an optimal sequential decision making strategy under this model therefore involves careful tradeoff of immediate and future rewards.

Here are some examples of possible policies. Suppose you decide that whenever the robot is in a standing position, you will make the robot walk and not run - this is a stationary Markovian (deterministic) policy. A more complex policy could be that whenever the robot is standing 2 or more steps away from the target location (5), in those states you will make it walk, otherwise you make it run. This is another Markovian stationary deterministic policy which is conservative in states farther from target and aggressive in states closer to the target. You can get

a randomized policy by making it walk, run or stay with some probability.In general you can change policies over time, it doesn't need to be stationary. May be initially you decided that you will always walk in standing state, but later on after realizing that you are moving very slowly, you changed your mind and started running in all states. In this case the agent is using different policies at different time steps, i.e., a nonstationary policy.

Example 2. Inventory control problem. Each month the manager of a warehouse determines current inventory (stock on hand) of a single product. Based on this information, she decides whether or not to order additional stock from a supplier. In doing so, she is faced with a trade-off between holding costs and the lost sales or penalties associated with being unable to satisfy customer demand for the product. The objective is to maximize some measure of profit over a given time horizon. Demand is a random variable with a probability distribution known to the manager. Let s_t denote the inventory on hand at the beginning of the tth time period, D_t be the random

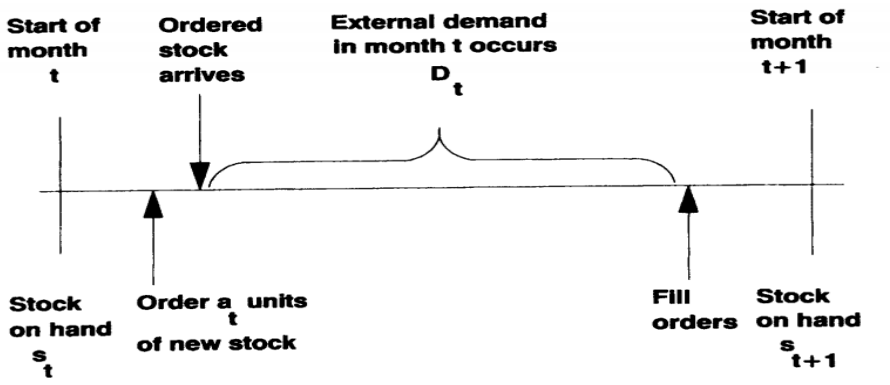


Figure 2: Timing of events in an inventory model (Figure taken from Puterman [1994].)

demand during this time period, and a_t be the number of units ordered by the inventory manager in the beginning of period t. We assume that the demand has a known time-homogeneous probability distribution $p_j = \Pr(D_t = j)$, $j = 0, 1, \ldots$. The inventory in the beginning of decision epoch $t+1$ referred to as s_{t+1}, is related to the inventory at decision epoch t, s_t, through the system equation

$$s_{t+1} = \max\{s_t + a_t - D_t, 0\} \equiv [s_t + a_t - D_t]^+.$$

That backlogging is not allowed implies the non-negativity of the inventory level. Denote by $O(u)$ the cost of ordering u units in any time period. Assuming a fixed cost K for placing orders and a variable cost $c(u)$ that increases with quantity ordered, we have

$$O(u) = [K + c(u)]1_{\{u>0\}}.$$

The cost of maintaining an inventory of u units for a time period is represented by a nondecreasing function $h(u)$. Finally, if the demand is j units and sufficient inventory is available to meet demand, the manager receives revenue with present value $f(j)$. In this model, the reward depends on the state of the system at the subsequent decision epoch, that is

$$r_t(s_t, a_t, s_{t+1}) = -O(a_t) - h(s_t + a_t) + f(s_t + a_t - s_{t+1}).$$

The goal of a inventory policy could be to maximize expected total reward in a finite horizon, or discounted reward if the firm cares more about near future.

Example 3. (running example) Here is another simple example of MDP model for a robot trying to walk. We eliminate the location and target location for the robot. The robot now just want to make as much progress as possible without falling. The robot can be in three states : Fallen state, Standing state, or Moving state. There are two possible actions: moving the robot legs slowly and moving the robot legs aggressively. Black arrows show what happens with slow action, and green arrows show what happens with the aggressive action. By slow action in

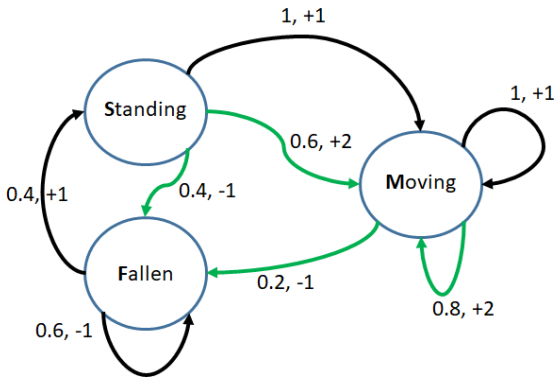


Figure 3: A simple MDP for the robot toy example

Fallen state, the robot may be able to stand up only with 0.4 prob to receive reward +1, but with 0.6 probability it may fall back and receive reward −1. The fast or aggressive action is not available in this state. In Standing state, slow action is very reliable and puts the robot in moving state with prob 1, also earning a reward 1. Fast action in a standing state can earn more reward when it is successful in transferring the robot to the moving state, but with 0.4 probability, it may make the robot fall, which means transfer to the Fallen state and a reward of −1. In moving state, again, the slow action is reliable, but fast action can earn more reward, with a risk of falling that is smaller than the risk in standing state.

Here, state space $S = \{'F', 'S', 'M'\}$, $A = \{'slow', 'fast'\}$. R is an $S \times A$ matrix and P is $S \times A \times S$ matrix.

$$R = \begin{bmatrix} (0.6 \times -1 + 0.4 \times 1) & 0 \\ 1 & (0.6 \times 2 + 0.4 \times -1) \\ 1 & (0.8 \times 2 + 0.2 \times -1) \end{bmatrix} = \begin{bmatrix} -0.2 & 0 \\ 1 & 0.8 \\ 1 & 1.4 \end{bmatrix}$$

$$P(s, \text{slow}, s') = \begin{bmatrix} 0.6 & 0.4 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad P(s, \text{fast}, s') = \begin{bmatrix} 1 & 0 & 0 \\ 0.4 & 0 & 0.6 \\ 0.2 & 0 & 0.8 \end{bmatrix}$$

4 Solving an MDP (Bellman equations)

4.1 Finite horizon

Bellman equations named after their discoverer Richard Bellman, provide a recursive formula for gain of an MDP. For finite horizon MDP this is simply dynamic programming.

Consider the toy example of robot trying to walk in Figure 3. Let the starting state is 'Standing'. Try to compute the optimal policy for horizon $H = 1, 2, 3, \ldots, 10$ for total reward criteria ($\gamma = 1$) *by enumeration*. For $H = 1$, the optimal policy simply maximizes immediate reward $\arg\max_a R(s, a)$ for $s =' Standing'$. And, therefore optimal policy is to take the slow action (black). For $H = 2$, the optimal policy involves deciding a two-stage decision. Deciding the first action (in 'Standing' state) involves enumerating the tree of all possible trajectories of state-action sequences starting from this state and every action. That is, $A^H(S)^{H-1}$ possibilities. A central idea in solving MDPs is that the Markovian structure can be used to make this computation tractable, using the simple idea of memoization (dynamic programming).

Bellman optimality equations. Let $V_k^*(s)$ be defined as the maximum total (discounted) reward achievable over a k length horizon starting in state s. Then,

$$V_k^*(s) = \max_\pi \mathbb{E}[\sum_{t=1}^{k} \gamma^{t-1} r_t | s_1 = s]$$

where maximum is taken over all (non-stationary) policies $\pi = (\pi_1, \ldots, \pi_k)$, $a_t = \pi_t(s_t), \mathbb{E}[r_t|s_t] = R(s_t, a_t), \Pr(s_{t+1} = s'|s_t, a_t) = P(s_t, a_t, s')$.

Then, we have optimal substructure property:

$$\begin{aligned}
V_k^*(s) &= \max_\pi \left\{ \mathbb{E}[r_1|s_1 = s] + \mathbb{E}[\mathbb{E}[\sum_{t=2}^{k} \gamma^{t-1} r_t | s_1 = s, s_2 = s']] \right\} \\
&= \max_a R(s,a) + \max_{\pi_2,\ldots,\pi_k} \sum_{s'} P(s,a,s') \mathbb{E}[\sum_{t=2}^{k} \gamma^{t-1} r_t | s_2 = s'] \\
&= \max_a R(s,a) + \gamma \sum_{s'} P(s,a,s') \left\{ \max_{\pi_1,\ldots,\pi_{k-1}} \mathbb{E}[\sum_{t=1}^{k-1} \gamma^{t-1} r_t | s_1 = s'] \right\} \\
&= \max_a R(s,a) + \gamma \sum_{s'} P(s,a,s') V_{k-1}^*(s'), k = 1, \ldots, H
\end{aligned}$$

And, by definition

$$\rho^*(s_1) = V_H^*(s_1)$$

This can be used to solve a finite horizon MDP by dynamic programming, by building a table of $H \times S$ values, starting from the last time step.

Example. Let's compute below for the toy example of robot MDP. Further examples are available in Section 4.6 of Puterman [1994].

Let's optimize for horizon $H = 4$. Now, $V_1^*(\cdot)$ is simply immediate reward maximization,

$$\begin{aligned}
V_1^*(F) &= 0(\text{fast action/do nothing}) \\
V_1^*(S) &= 1(\text{slow action}) \\
V_1^*(M) &= 1.4(\text{fast action})
\end{aligned}$$

This suggest that if time horizon is 1, the robot should not try to get up from fallen state.

$$\begin{aligned}
V_2^*(F) &= \max\{-0.2 + 0.4 \times 1 + 0.6 \times 0, 0 + 0\} = 0.2(\text{slow action}) \\
V_2^*(S) &= \max\{1 + 1.4, 0.8 + 0.6 \times 1.4 + 0.4 \times 0\} = 2.4(\text{slow action}) \\
V_2^*(M) &= \max\{1 + 1.4, 1.4 + 0.8 \times 1.4 + 0.2 \times 0\} = 2.56(\text{fast action})
\end{aligned}$$

$$\begin{aligned}
V_3^*(F) &= \max\{-0.2 + 0.4 \times 2.4 + 0.6 \times 0.2, 0 + 0.2\} = 0.88(\text{slow action}) \\
V_3^*(S) &= \max\{1 + 2.56, 0.8 + 0.6 \times 2.56 + 0.4 \times 0.2\} = 3.56(\text{slow action}) \\
V_3^*(M) &= \max\{1 + 2.56, 1.4 + 0.8 \times 2.56 + 0.2 \times 0.2\} = \max\{3.56, 3.488\} = 3.56(\text{slow action})
\end{aligned}$$

(If you use $\gamma < 1$, it might take more time steps for the action in state M to become slow action, depending on how small γ is. Intuitively, if horizon is short or future is either discounted heavily you might want to be more aggressive).

In the next iteration, the policy remains the same:

$$\begin{aligned}
V_4^*(F) &= \max\{-0.2 + 0.4 \times 3.56 + 0.6 \times 0.88, 0 + 0.88\} = \max\{1.752, 0.88\} = 1.752(\text{slow action}) \\
V_4^*(S) &= \max\{1 + 3.56, 0.8 + 0.6 \times 3.56 + 0.4 \times 0.88\} = \max\{4.56, 3.24\} = 4.56(\text{slow action}) \\
V_4^*(M) &= \max\{1 + 3.56, 1.4 + 0.8 \times 3.56 + 0.2 \times 0.88\} = \max\{4.56, 4.4\} = 4.56(\text{slow action})
\end{aligned}$$

4.2 Infinite horizon discounted reward

Henceforth we will assume finite and discrete state space S, finite and discrete action space A, bounded rewards $R(s,a)$ and discount $\gamma < 1$. In this case, there exists an optimal stationary policy. We abuse notation, to denote a stationary policy $(\pi, \pi, \pi, \ldots)$, as π. Therefore, we are effectively looking for a stationary policy $\pi^* \in \arg\max \rho^\pi(s_1)$.

Value of a policy π in a given state s at time t is the gain when starting from state s.

$$V_\gamma^\pi(s) = \lim_{T\to\infty} \mathbb{E}[\sum_{t=1}^{T} \gamma^{t-1} r_t | s_1 = s], \forall s.$$

Note that gain of a policy is simply $\rho^\pi(s_1) = V^\pi(s_1)$, i.e., the value from the starting state. (Value is also referred to as 'cost-to-go' when cost-based version of MDP is considered. In that version, instead of reward, you observe a cost, and the goal is to minimize total/average/discounted cost).

Bellman equations for value of a policy. In infinite horizon case, the value of policy only depends on the state and not the time, and satisfies the following recursive relation.

$$V_\gamma^\pi(s) = \mathbb{E}_{a\sim\pi(s), s'\sim P(s,a)} \left[R(s,a,s') + \gamma V^\pi(s') \right], \text{ or,}$$
$$V_\gamma^\pi = \mathbf{R}^\pi + \gamma P^\pi V^\pi$$

Proof:

$$\begin{aligned} V_\gamma^\pi(s) &= \mathbb{E}[r_1 + \gamma r_2 + \gamma^2 r_3 + \gamma^3 r_4 + \ldots | s_1 = s] \\ &= E[r_1 | s_1 = s] + \gamma \mathbb{E}[\mathbb{E}[r_2 + \gamma r_3 + \gamma^2 r_4 + \ldots | s_2] | s_1 = s] \end{aligned}$$

The first term here is simply the expected reward in state s when action is given by $\pi(s)$. The second term is γ times the value function at $s_2 \sim P(s, \pi(s), \cdot)$

$$\begin{aligned} V_\gamma^\pi(s) &= \mathbb{E}[R(s,\pi(s),s_1) + \gamma V_\gamma^\pi(s_2) | s_1 = s] \\ &= R(s,\pi(s)) + \gamma \sum_{s_2 \in S} P(s,\pi(s),s_2) V_\gamma^\pi(s_2) \\ &= R^\pi(s) + \gamma [P^\pi V_\gamma^\pi](s) \end{aligned}$$

Bellman optimality equations. Let $V_\gamma^*(s) = \max_\pi V_\gamma^\pi(s)$.

$$V_\gamma^*(s) = \max_a R(s,a) + \gamma \sum_{s'} P(s,a,s') V_\gamma^*(s')$$

And, by definition

$$\rho^*(s) = V^*(s)$$

Proof: for all s, from the theorem ensuring stationary optimal policy:

$$\begin{aligned} V_\gamma^*(s) = \max_\pi V_\gamma^\pi(s) &= \max_\pi \mathbb{E}_{a\sim\pi(s), s'\sim P(s,a)} \left[R(s,a,s') + \gamma V_\gamma^\pi(s') \right] \\ &\leq \max_a R(s,a) + \gamma \sum_{s'} P(s,a,s') \max_\pi V_\gamma^\pi(s') \\ &= \max_a R(s,a) + \gamma \sum_{s'} P(s,a,s') V_\gamma^*(s') \end{aligned}$$

Now, if the above inequality is strict then the value of state s can be improved by using a (possibly non-stationary) policy that uses action $\arg\max_a R(s,a)$ in the first step. This is a contradiction to the definition $V_\gamma^*(s)$. Therefore,

$$V_\gamma^*(s) = \max_a R(s,a) + \gamma \sum_{s'} P(s,a,s') V_\gamma^*(s')$$

Technically, above only shows that V_γ^* satisfies the Bellman equations. Theorem 6.2.2 (c) in Puterman [1994] shows that V^* is in fact unique solution of above equations. Therefore, satisfying these equations is sufficient to guarantee optimality, so that it is not difficult to see that the deterministic (stationary) policy

$$\pi_\gamma^*(s) = \arg\max_a R(s,a) + \gamma \sum_{s'} P(s,a,s') V_\gamma^*(s')$$

is optimal (see Puterman [1994] Theorem 6.2.7 for formal proof).

And, by Bellman optimality equations, $V^*_\gamma = R^{\pi^*} + \gamma P^{\pi^*} V^*_\gamma$, i.e., $V^*_\gamma = (I - \gamma P^{\pi^*})^{-1} R^{\pi^*}$, where the inverse exists for $\gamma < 1$.

Linear programming. The fixed point for above Bellman optimality equations can be found by formulating a linear program. It amounts to :

$$\begin{aligned} \min_{\mathbf{v} \in \mathbb{R}^S} \quad & \textstyle\sum_s v_s \\ \text{subject to} \quad & v_s \geq R(s,a) + \gamma P(s,a)^\top \mathbf{v} \quad \forall a, s \end{aligned}$$

Proof. Let V^* satisfies the Bellman optimality equations. First we show that V^* will be a feasible and optimal solution of the above LP. V^* clearly satisfies the constraints of the above LP. Next, we show that V^* minimizes the objective of the LP. Consider any feasible $\mathbf{v}$. Then,

$v_s \geq R(s,a) + P(s,a)^\top \mathbf{v}, \forall s, a$ implies that

$$v_s \geq R(s, \pi^*(s)) + \gamma P(s, \pi^*(s))^\top \mathbf{v}, \forall s$$

(Above is written assuming optimal policy π^* is deterministic, which is in fact true in the infinite horizon discounted reward case.) Or,

$$(I - \gamma P^{\pi^*}) \mathbf{v} \geq R^{\pi^*}$$

Because $\gamma < 1$, $(I - \gamma P^\pi)^{-1}$ exists for all π, and for any $u \geq 0$

$$(I - \gamma P^\pi)^{-1} u = (I + \gamma P^\pi + \gamma^2 (P^\pi)^2 + \cdots) u \geq u$$

Therefore, from above

$$(I - \gamma P^{\pi^*})^{-1} ((I - \gamma P^{\pi^*}) \mathbf{v} - R^{\pi^*}) \geq 0$$

Or,

$$\mathbf{v} \geq (I - \gamma P^{\pi^*})^{-1} R^{\pi^*} = V^*$$

Therefore, $\mathbf{w}^\top \mathbf{v} \geq \mathbf{w}^\top V^*$ for $\mathbf{w} > 0$.

For the other direction, consider any optimal solution $\mathbf{v}^*$ of the above LP. Then, we show that it satisfies Bellman equation. It is easy to see from the constraints that it satisfies for all s,

$$v^*_s \geq \max_a R(s,a) + \gamma P(s,a)^T \mathbf{v}^*$$

Now since the objective minimizes $\sum_s v_s$, and since there exists a feasible solution that satisfies above constraints with inequality (optimal value V^* is such a solution) it must be the case that for optimal v^*,

$$v^*_s = \max_a R(s,a) + \gamma P(s,a)^T \mathbf{v}^*$$

Therefore, the optimal solution is a fixed point of the Bellman equations.

4.3 Infinite horizon average reward

Gain of a policy. Gain of a policy in this case is asymptotic average reward starting from state s,

$$\rho^\pi(s) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}[\sum_{t=1}^{T} r_t | s_1 = s]$$

This is related to finite time undiscounted value, and infinte horizon discounted value, in the following ways:

- $$\rho^\pi(s) = \lim_{T \to \infty} \frac{1}{T} V^\pi_T(s)$$

  where $V^\pi_T(s) = \mathbb{E}[\sum_{t=1}^{T} r_t | s_1 = s]$.

- $$\rho^\pi(s) = \lim_{\gamma \to 1}(1-\gamma)V_\gamma^\pi(s)$$

  For an intuitive explanation of above relation consider $\gamma = 1 - (1/T)$ with T going to infinity.

The above connection to discounted value function is is formally established using Laurent series expansion of $V_\gamma^\pi = (I - \gamma P^\pi)^{-1}\mathbf{R}^\pi$, which is given by

$$V_\gamma^\pi = \frac{1}{(1-\gamma)}\rho^\pi + h^\pi + f(\gamma)$$

where $f(\gamma)$ denotes a vector that goes to 0 has $\gamma \to 1$. For further details, see Section 8.2.2 of [Puterman, 1994].

Bias of a policy. For average reward case, an important quantity is the total deviation of reward from asymptotic average reward. This is referred to as the **'bias'** of a policy. Bias of a policy π in state s is given by

$$h^\pi(s) = \lim_{T\to\infty} \mathbb{E}[\sum_{t=1}^{T}(r_t - \rho^\pi(s_t))|s_1 = s] = \lim_{T\to\infty} \mathbb{E}[\sum_{t=1}^{T}(R(s_t, \pi(s_t)) - \rho^\pi(s_t))|s_1 = s].$$

The limit in above is Cesaro limit, which exists, more details in Section 8.2 of Puterman [1994].

A connection between value finite time value ($V_T^\pi(s) = \mathbb{E}[\sum_{t=1}^{T} r_t|s_1 = s]$) and bias is that: if two states s, s' are in the same irreducible class (i.e., can be reached from each other in finite expected time, under policy π) then

$$h^\pi(s) - h^\pi(s') = \lim_{T\to\infty}(V_T^\pi(s) - V_T^\pi(s'))$$

And, connection to discounted value function:

$$h^\pi(s) - h^\pi(s') = \lim_{\gamma\to 1}(V_\gamma^\pi(s) - V_\gamma^\pi(s'))$$

Remark: To see why we need the two states to be in the same irreducible class, note that being in the same irreducible class ensures that $\rho^\pi(s_t) = \rho^\pi(s_t')$ for all the states visited from s vs s'.

4.3.1 Bellman equations for evaluating a policy.

Assume a policy π is such that all the states reached form a single recurrent class. Then, $\rho^\pi(s) = \rho^\pi(s'), \forall s, s'$, and

$$h^\pi(s) + \rho^\pi(s) = \mathbb{E}_{a\sim\pi(s), s'\sim P(s,a)}\left[R(s,a,s') + h^\pi(s')\right], \forall s$$

Or, in compact notation:

$$h^\pi + \rho^\pi = \mathbf{R}_\pi + P^\pi h^\pi$$

Also, for any $h \in \mathbb{R}^n, \rho \in \mathbb{R}$ that satisfy the equations,

$$h + \rho\mathbf{e} = \mathbf{R}_\pi + P^\pi h$$

we have that $\rho = \rho^\pi$ and $h = h^\pi + c\mathbf{e}$ for some constant c. Here $\mathbf{e}$ is the S-dimensional vector of all 1s.

Proof. (assumes finite or countable state space and policy space, and deterministic policy π for simplicity. Similar derivation can be done for randomized policy with notational changes.)

First we show that h^π, ρ^π satisfy the Bellman evaluation equations. We can use the Bellman equations for discounted value:

$$V_\gamma^\pi = \mathbf{R}_\pi + \gamma P^\pi V_\gamma^\pi$$

Subtract γV_γ^π from both sides:

$$V_\gamma^\pi(1-\gamma) = \mathbf{R}_\pi + \gamma P^\pi V_\gamma^\pi - \gamma V_\gamma^\pi$$

For state s:

$$\begin{aligned} V_\gamma^\pi(s)(1-\gamma) &= \mathbf{R}_\pi(s) + \sum_{s'} \gamma P^\pi(s,s') V_\gamma^\pi(s') - \gamma V_\gamma^\pi(s) \\ &= \mathbf{R}_\pi(s) + \sum_{s'} \gamma P^\pi(s,s')(V_\gamma^\pi(s') - V_\gamma^\pi(s)) \\ \lim_{\gamma\to 1} V_\gamma^\pi(s)(1-\gamma) &= \mathbf{R}_\pi(s) + \sum_{s'} P^\pi(s,s') \lim_{\gamma\to 1} \gamma(V_\gamma^\pi(s') - V_\gamma^\pi(s)) \\ \rho^\pi(s) &= \mathbf{R}_\pi(s) + \sum_{s'} P^\pi(s,s')(h^\pi(s') - h^\pi(s)) \end{aligned}$$

Therefore,

$$\rho^\pi = \mathbf{R}_\pi + P^\pi h^\pi - h^\pi$$

This shows that ρ^π, h^π satisfies Bellman equations . For a proof of that second part, we want to show that any feasible ρ, h to the equations

$$\rho e + h = R^\pi + P^\pi h$$

gives gain, bias of policy π.

To see this (rough proof) multiply by $(P^\pi)^i$ on both sides and sum for $i = 0, 1, 2, \ldots, T-1$. Then,

$$T\rho e + \sum_{i=0}^{T-1} (P^\pi)^i h = \sum_{i=0}^{T-1} (P^\pi)^i R^\pi + \sum_{i=0}^{T-1} (P^\pi)^{i+1} h$$

which is equivalent to

$$T\rho e + h = \sum_{i=0}^{T-1} (P^\pi)^i R^\pi + (P^\pi)^T h$$

Dividing by T and taking $T \to \infty$, we get

$$\rho e = \lim_{T\to\infty} \frac{1}{T} \sum_{i=0}^{T-1} (P^\pi)^i R^\pi$$

That is ρ is the average reward of policy π.

And rearranging the terms and taking $T \to \infty$, we get

$$h - \lim_{T\to\infty} (P^\pi)^T h = \lim_{T\to\infty} (\sum_{i=0}^{T-1} (P^\pi)^i R^\pi - \rho e)$$

which (under certain technical conditions) gives

$$h + ce = \lim_{T\to\infty} \sum_{i=0}^{T-1} ((P^\pi)^i R^\pi - \rho e)$$

for a constant c, i.e., h is the bias of policy π within a constant c. For a more rigorous proof, refer to Theorem 8.2.6 in Puterman [1994]. □

Example. For MDP in Figure 3, the consider the policy (say π) that takes the slow action in all states. The gain of this policy is $\rho^\pi(s) = 1$ for all states $s \in \{F, S, M\}$. The bias of this policy is $h^\pi(F) = (-0.2 - 1) \times (1/0.4) = -3, h^\pi(S) = 0, h^\pi(M) = 0$. (For calculating $h^\pi(F)$, note that the expected number of steps spent in Fallen state when taking slow actions is $1/0.4$, after that the reward of the given policy is 1).

Check that the bias and gain of this policy satisfy the Bellman equations stated above.

$$\underbrace{\begin{bmatrix} -3 \\ 0 \\ 0 \end{bmatrix}}_{h^\pi} + \underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}}_{\rho^\pi \mathbf{e}} = \underbrace{\begin{bmatrix} -0.2 \\ 1 \\ 1 \end{bmatrix}}_{R^\pi} + \underbrace{\begin{bmatrix} 0.6 & 0.4 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}}_{P^\pi} \underbrace{\begin{bmatrix} -3 \\ 0 \\ 0 \end{bmatrix}}_{h^\pi}$$

4.3.2 Bellman optimality equations.

We will make an additional 'communicating MDP' assumption.

Definition 4. *An MDP is called* **communicating** *if for any two states s, s', there exists a policy such that the expected number of steps to reach s' from s is finite.*

A convenient fact is that optimal gain does not depend on the starting state for such MDPs.

Theorem 5 (Puterman [1994], Theorem 8.3.2 in Section 8.3.3)**.** *For communicating MDP, for optimal gain policy $\rho^*(s) = \rho^*$, i.e., optimal average infinite horizon reward does not depend on starting state.*

Proof. (Sketch) Suppose that that there exists $s_1 \neq s_2$ such that $\rho^*(s_1) > \rho^*(s_2)$. Since the MDP is communicating there exists a policy π_0 using which we can go from s_2 to s_1 in time τ with finite expected value, say $\mathbb{E}[\tau] \leq D$. Then we can construct a (possibly non-stationary) policy, which first goes from s_2 to s_1 using π_0 in at most D steps in expectation, and then uses the optimal policy (say π_1) for s_1. Such a policy will have infinite horizon average reward $\rho^*(s_1)$ which is strictly greater than $\rho^*(s_2)$, thus violating the optimality of $\rho^*(s_2)$. □

Recall stationary (possibly non-deterministic) optimal policy π^* exists for this setting. And, $\rho^{\pi^*}(s_1) = \rho^*$ is independent of the starting state.

Lemma 6 (Bellman optimality equations for average reward MDP)**.** *For any communicating MDP, Bellman optimality equations state that optimal gain is $\rho^* = \rho$ where (ρ, h) is a feasible solution to the following equations:*

$$\rho + h(s) = \max_a R(s,a) + \sum_{s'} P(s,a,s')h(s'), \forall s$$

Also, for any feasible solution (ρ, h) to the Bellman equations, we can get an optimal policy π^* defined as

$$\pi^*(s) \in \arg\max_a R(s,a) + P(s,a)^\top h,$$

with $\rho = \rho^{\pi^*}$ and $h = h^{\pi^*} + c\mathbf{e}$ for some constant c.

Proof. We prove the above lemma partially. We assume that a solution $(\rho^*, h^*)$to the above equations exist. Then, we show that $\rho^* \geq \rho^\pi$ for every policy $\pi = (\pi_1, \ldots, \pi_T, \ldots,)$ with equality achieved by the arg max policy. Proof of existence of such a solution is more intricate, and is shown by using relation with discounted model through Laurent series expansion (refer to Section 9.1.3 of Puterman [1994]).

Now, using the equation for first step policy as π_1 we have:

$$\rho^*\mathbf{e} \geq R_{\pi_1} + (P_{\pi_1} - I)h^*$$

Using the equation for π_2, and multiplying by P_{π_1} on both sides

$$\rho^*\mathbf{e} \geq P_{\pi_1}R_{\pi_2} + P_{\pi_1}(P_{\pi_2} - I)h^*$$

Similarly, for any $t = 1, 2, \ldots,$ we can get

$$\rho^*\mathbf{e} \geq P_{\pi_1}P_{\pi_2}\cdots P_{\pi_{t-1}}R_{\pi_t} + P_{\pi_1}P_{\pi_2}\cdots P_{\pi_{t-1}}(P_{\pi_t} - I)h^*$$

On adding above equations for for $t = 1, \ldots, T$, the first term on the right hand side adds to value of policy in T rounds. The second term reduces to $(\prod_{i=1}^{T} P_{\pi_i} - I)h^*$. Therefore,

$$T\rho^*\mathbf{e} \geq V_T^\pi + (\prod_{i=1}^{T} P_{\pi_i} - I)h^*$$

Dividing by T and taking limit $T \to \infty$:

$$\rho^*\mathbf{e} \geq \lim_{T\to\infty} \frac{1}{T}V_T^\pi + \frac{1}{T}(\prod_{i=1}^{T} P_{\pi_i} - I)h^*$$

we get

$$\rho^* \mathbf{e} \geq \rho^\pi$$

This proves the first part of the lemma.

Now, for any arg max policy π, we have

$$\rho^* \mathbf{e} = R_\pi + (P_\pi - I)h^*$$

Therefore, ρ^*, h^* satisfy the Bellman evaluation equations for policy π, therefore it follows from the previous subsection that $\rho^* = \rho^\pi$, and $h^* = h^\pi + ce$ □

Linear programming. Based on above discussion, for infinite horizon average reward case, the fixed point for Bellman optimality equations can be found by formulating a linear program (assuming communicating MDP).

$$\begin{aligned} \min_{\rho \in R, \mathbf{v} \in \mathbb{R}^S} \quad & \rho \\ \text{subject to} \quad & \rho \geq R(s,a) + P(s,a)^\top \mathbf{h} - h_s \quad \forall a, s \end{aligned}$$

If the MDP is multichain and not necessarily communicating, the optimality equations can still be formulated, but they are slightly more complex. Interested readers may refer to Chapter 9.1 of Puterman [1994].

However, solving LP is slow, and therefore, faster iterative methods are used.

5 Iterative algorithms (discounted reward case)

Using dynamic programming directly may not be very efficient especially for large/infinite horizon case. Below, we discuss some popular iterative methods that are more efficient than linear programming. For succinctness, we limit our discussion primarily to the discounted reward case. Similar algorithms and convergence results are available for the average reward case. For the average reward case, more conditions on the transition matrix are required for convergence. See Chapter 8 of Puterman [1994] for more details.

5.1 Value Iteration.

Value iteration iteration computes optimal value function (value vector $\mathbf{v}$ in above), not the explicit policy. A near optimal policy is found at the end of the algorithm as the greedy policy based on the computed value functions.

Pseudocode

1. Start with an arbitrary initialization $\mathbf{v}^0$. Specify $\epsilon > 0$
2. **Repeat** for $k = 1, 2, \ldots$ **until** $\|\mathbf{v}^k - \mathbf{v}^{k-1}\|_\infty \leq \epsilon \frac{(1-\gamma)}{2\gamma}$:
   - for every $s \in S$, improve the value vector as:

$$\mathbf{v}^k(s) = \max_{a \in A} R(s,a) + \gamma \sum_{s'} P(s,a,s') v^{k-1}(s'), \tag{1}$$

3. Compute a near-optimal policy as

$$\pi(s) \in \arg\max_a R(s,a) + \gamma P(s,a)^\top \mathbf{v}^k \tag{2}$$

Bellman operator It is useful to represent the iterative step (1) using operator $L : \mathbb{R}^S \to \mathbb{R}^S$.

$$\begin{aligned} LV(s) &:= \max_{a\in A} R(s,a) + \gamma \sum_{s'} P(s,a,s')V(s') \\ L^\pi V(s) &:= \mathbb{E}_{a\in\pi(s)}[R(s,a) + \gamma \sum_{s'} P(s,a,s')V(s')] \end{aligned} \tag{3}$$

Then, (1) is same as

$$\mathbf{v}^k = L\mathbf{v}^{k-1} \tag{4}$$

Also, for any policy π, if V^π denotes its value function, then, by Bellman equations:

$$V^* = LV^*, V^\pi = L^\pi V^\pi \tag{5}$$

Below is a useful 'contraction' property of this operator, which underlies the convergence properties of all DP based iterative algorithms.

Lemma 7. *The operator $L(\cdot)$ and $L^\pi(\cdot)$ defined by* (3) *are contraction mappings, i.e.,*

$$\|Lv - Lu\|_\infty \le \gamma \|v - u\|_\infty.$$

$$\|L^\pi v - L^\pi u\|_\infty \le \gamma \|v - u\|_\infty.$$

Proof. First assume $Lv(s) \ge Lu(s)$. Let $a_s^* = \arg\max_{a\in A} R(s,a) + \gamma \sum_{s'} P(s,a,s')v(s')$

$$\begin{aligned} 0 &\le Lv(s) - Lu(s) \\ &\le R(s,a_s^*) + \gamma \sum_{s'} P(s,a_s^*,s')v(s') - R(s,a_s^*) - \gamma \sum_{s'} P(s,a_s^*,s')u(s') \\ &= \gamma P(s,a_s^*)^\top (v-u) \\ &\le \gamma \|v-u\|_\infty \end{aligned}$$

Repeating a symmetric argument for the case $Lu(s) \ge Lv(s)$ gives the lemma statement. Similar proof holds for L^π. □

Convergence

Theorem 8 (Theorem 6.3.3, Section 6.3.2 in Puterman [1994])**.** *The convergence rate of the above algorithm is linear at rate γ. Specifically,*

$$\|\mathbf{v}^k - V^*\|_\infty \le \frac{\gamma^k}{1-\gamma} \|v^1 - v^0\|_\infty$$

Further, let π^k be the policy given by (2) *using v^k. Then,*

$$\|V^{\pi^k} - V^*\|_\infty \le \frac{2\gamma^k}{1-\gamma} \|v^1 - v^0\|_\infty$$

Proof. By Bellman equations $V^* = LV^*$.

$$\begin{aligned} \|V^* - v^k\|_\infty &= \|LV^* - v^k\|_\infty \\ &\le \|LV^* - Lv^k\|_\infty + \|Lv^k - v^k\|_\infty \\ &= \|LV^* - Lv^k\|_\infty + \|Lv^k - Lv^{k-1}\|_\infty \\ &\le \gamma \|V^* - v^k\| + \gamma \|v^k - v^{k-1}\| \\ &\le \gamma \|V^* - v^k\| + \gamma^k \|v^1 - v^0\| \\ \|V^* - v^k\|_\infty &\le \frac{\gamma^k}{1-\gamma} \|v^1 - v^0\| \end{aligned}$$

Let $\pi = \pi^k$ be the policy at the end of k iterations. Then, $V^\pi = L^\pi V^\pi$ by Bellman equations. Further, by definition of $\pi = \pi^k$,

$$L^\pi v^k(s) = \max_a R(s,a) + \gamma \sum_{s'} P(s,a,s') v^k(s') = Lv^k(s).$$

Therefore,

$$\begin{aligned}
\|V^\pi - v^k\|_\infty &= \|L^\pi V^\pi - v^k\|_\infty \\
&\leq \|L^\pi V^\pi - L^\pi v^k\|_\infty + \|L^\pi v^k - v^k\|_\infty \\
&= \|L^\pi V^\pi - L^\pi v^k\|_\infty + \|Lv^k - Lv^{k-1}\|_\infty \\
&\leq \gamma \|V^\pi - v^k\| + \gamma \|v^k - v^{k-1}\| \\
\|V^\pi - v^k\|_\infty &\leq \frac{\gamma}{1-\gamma} \|v^k - v^{k-1}\| \\
&\leq \frac{\gamma^k}{1-\gamma} \|v^1 - v^0\|
\end{aligned}$$

Adding the two results above:

$$\|V^\pi - V^*\|_\infty \leq \frac{2\gamma^k}{1-\gamma} \|v^1 - v^0\|_\infty$$

□

In average reward case, the algorithm is similar, but the Bellman operator used to update the values is now $LV(s) = \max_a r_{s,a} + P(s,a)^\top V$. Also, here $\mathbf{v}^k$ will converge to $\mathbf{v}^* + c\mathbf{e}$ for some constant c. Therefore, the stopping condition used is instead $sp(v^k - v^{k-1}) \leq \epsilon$ where $\text{sp}(v) := \max_s v_s - \min_s v_s$. That is, span is used instead of L_∞ norm. Further since there is no discount ($\gamma = 1$), a condition on the transition matrix is required to prove convergence. Let

$$\gamma := \max_{s,s',a,a'} 1 - \sum_{j \in S} \min\{P(s,a,j), P(s',a',j)\}$$

Then, linear convergence with rate γ is guaranteed if $\gamma < 1$. Or in other words, if

$$\min_{s,s',a,a'} \sum_{j \in S} \min(P(s,a,j), P(s',a',j)) > 0,$$

i.e., on taking two different actions in any two different states, there is a positive probability to reach an identical state. This condition ensures that the Bellman operator in this case: is still a contraction - although in terms of span, i.e., $span(Lv - Lu) \leq \gamma span(v - u)$. For more details, refer to Section 8.5.2 in Puterman [1994].

5.2 Q-value iteration

A variation of value iteration is obtained by updating Q-values instead. This variation will be useful for deriving Q-learning algorithm for learning optimal policies from sample observations.

Define $Q^*(s,a)$ as the expected utility on taking action a in state s, and thereafter acting optimally. Then, $V^*(s) = \max_a Q^*(s,a)$. Therefore, Bellman equations can be written as,

$$Q^*(s,a) = R(s,a) + \gamma \sum_{s'} P(s,a,s') \left(\max_{a'} Q^*(s',a') \right)$$

Similarly, we define $Q^\pi(s,a)$ as the expected utility on taking action a in state s, and thereafter using policy π. So that $V^\pi(s) = \mathbb{E}_{s \sim \pi(s)}[Q^\pi(s,a)]$

$$Q^\pi(s,a) = R(s,a) + \gamma \sum_{s'} P(s,a,s') \mathbb{E}_{a' \sim \pi(s')}[Q^\pi(s',a')]$$

Based on above, the Q-value-iteration algorithm for finding optimal policy can be derived as follows:

Pseudocode

1. Start with an arbitrary initialization $\mathbf{Q}^0 \in \mathbb{R}^{S \times A}$.
2. In every iteration k, improve the Q-value vector as:

$$\mathbf{Q}^k(s,a) = R(s,a) + \gamma \sum_{s'} P(s,a,s') \left(\max_{a'} Q^{k-1}(s',a') \right), \forall s, a$$

3. Stop if $\|Q^k - Q^{k-1}\|_\infty$ is small.

5.3 Policy iteration.

Direct method that finds optimal policy.

Pseudocode

1. Start with an arbitrary initialization of policy π^0.
2. In every iteration $k = 0, 1, \ldots,$,
   - (Policy evaluation) Compute $V^{\pi^k} = (I - \gamma P^{\pi^k})^{-1} R^\pi$, the value of policy π^k.
   - (Greedy Policy improvement) Compute new policy

$$\pi^{k+1}(s) := \arg\max_a R(s,a) + \gamma \sum_{s'} P(s,a,s') V^{\pi^k}(s'), \forall s$$

3. Stop when $\pi^{k+1} = \pi^k$.

Relaxed stopping criteria may include stopping after a certain number of iterations given by the error bounds discussed below, or when $\|V^{\pi^{k+1}} - V^{\pi^k}\|_\infty$ is small.

For computing value of a policy π, one can also use an iterative procedure like value iteration, where we start from a v^0 vector and update $v^i = L^\pi v^{i-1}$ in each iteration i.

Note that the Greedy Policy can be equivalently written as:

$$\pi^{k+1}(s) := \arg\max_a Q^{\pi^k}(s,a), \forall s$$

Q-values of a policy can be learned using an iterative procedure like the Q-value iteration. This way of rewriting the algorithm will be useful for deriving the learning version of this algorithm when the reward and transition model R and P are unknown and Q-values will be learned from sample observations.

Geometric convergence. Similar to the value iteration method, the geometric convergence holds for policy iteration.

Theorem 9. *For the policy π^k computed in the k^{th} iteration of policy iteration, we have*

$$\|V^{\pi^k} - V^*\|_\infty \leq \gamma^k \|V^{\pi_0} - V^*\|_\infty$$

The following step quantifies the improvement in the policy obtained from the greedy policy improvement step.

Lemma 10.

$$V^{\pi^k} \leq L V^{\pi^k} \leq V^{\pi^{k+1}}$$

Proof. Let π be any policy, and π' be the policy obtained by the greedy step, i.e.,

$$\pi'(s) = \arg\max_a R(s,a) + \gamma \sum_{s'} P(s,a,s')V^{\pi}(s'), \forall s$$

Then, by definition of L and $L^{\pi'}$, we have,

$$LV^{\pi} = L^{\pi'}V^{\pi}$$

Also, by Bellman equations,

$$V^{\pi} = L^{\pi}V^{\pi} \leq LV^{\pi}$$

Combining the last two inequalities, we have

$$V^{\pi} \leq LV^{\pi} = L^{\pi'}V^{\pi} \tag{6}$$

Using above as base case, we prove by induction that for all $i = 1, 2, 3, \ldots,$

$$V^{\pi} \leq LV^{\pi} \leq (L^{\pi'})^i V^{\pi} \tag{7}$$

Suppose that above statement is true for i, i.e.,

$$V^{\pi} \leq (L^{\pi'})^i V^{\pi}$$

Then, using monotonicity of $L^{\pi'}$, If we apply $L^{\pi'}$ on both sides:

$$L^{\pi'}V^{\pi} \leq (L^{\pi'})^{i+1} V^{\pi}$$

Plugging this in (6),

$$V^{\pi} \leq LV^{\pi} = L^{\pi'}V^{\pi} \leq (L^{\pi'})^{i+1} V^{\pi}$$

which proves statement (7) by induction.

Now, taking i to ∞ in (7), since $(L^{\pi'})^i V^{\pi}$ will converge to $V^{\pi'}$ (this follows from convergence of value iteration for evaluating a policy)

$$V^{\pi} \leq LV^{\pi} \leq V^{\pi'}$$

□

Proof of Theorem 9. From the previous lemma, we have

$$V^{\pi^k} \geq LV^{\pi^{k-1}}$$

Subtracting from V^* and using $V^* = LV^*$,

$$V^* - V^{\pi^k} \leq LV^* - LV^{\pi^{k-1}}$$

so that

$$\|V^* - V^{\pi^k}\|_\infty \leq \|LV^* - LV^{\pi^{k-1}}\|_\infty \leq \gamma \|V^* - V^{\pi^{k-1}}\|_\infty$$

Applying this repeatedly for $k-1, \ldots, 1$, we get

$$\|V^* - V^{\pi^k}\|_\infty \leq \gamma^k \|V^* - V^{\pi^0}\|_\infty$$

□

5.4 Exercise

Use policy iteration and value iteration to compute optimal policy for the MDP in Figure 3 by hand.

6 Reinforcement learning algorithms

Reinforcement learning is essentially the sequential decision problem when the underlying MDP model (state transition probabilities and reward function) is either unknown or too difficult (large) to solve. We have seen some algorithms (value iteration, policy iteration, linear programming) for solving MDPs. There are two main challenges in using those for reinforcement learning problems:

- The model: $R(s,a), P(s,a,s')$ is not available in reinforcement learning. The model however may be accessible as a blackbox to generate samples. The challenge for RL algorithms is to (implicitly) learn this model from samples, while computing optimal policy. It is therefore important to consider both sample complexity and computation complexity when designing these algorithms.
- Number of states in most RL problems is too large for *tabular* methods like those discussed before to be scalable.

'Modern reinforcement learning' broadly refers to the algorithmic approaches to tackle these challenges, in order to solve a large/unknown MDP by learning and approximation. The focus is largely on scaling up reinforcement learning to make it possible to find complex effective policies for realistic tasks. The algorithmic approaches can be categorized as follows.

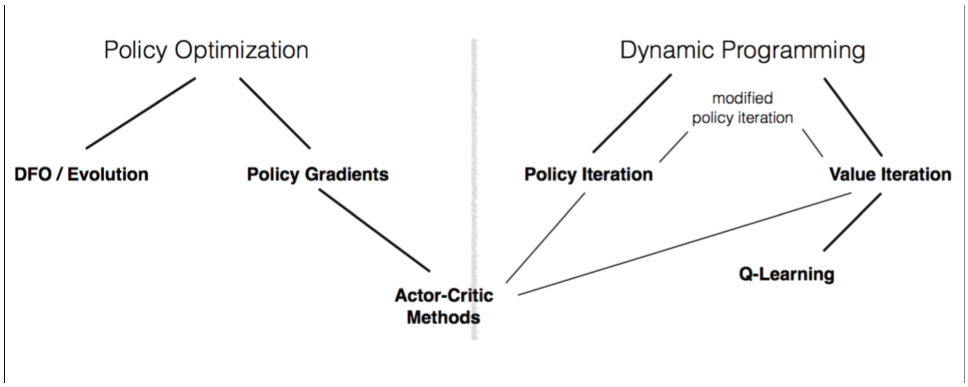


Figure 4: Direct Learning algorithms for RL (Drawing from Pieter Abbeel's slides)

- Direct Learning: Here, the optimal control policy is learned without first learning an explicit model. Such schemes include:
  - Approximate dynamic programming based approaches (Q-learning, TD learning)
  - Direct Policy search (policy gradient, genetic algorithms)
  - and their combinations (Actor-critic)
- Indirect Learning or Model based RL: Estimate an explicit model of the environment, and compute an optimal policy for the estimated model (e.g., Certainty Equivalence and R-MAX). This means learning estimates ($\hat{P}$ and $\hat{R}$) when the model is small but unknown. When the model is large, approximate compact representations of the model are learned. This approach provides a good framework for incorporating prior knowledge about the model into RL algorithms.

A further distinction in the literature is based on the performance criteria: the key difference between the two criteria listed below is whether we care about the performance of the policy found at the end of the algorithm (PAC analysis) or the performance during the execution of the algorithm (regret analysis).

- PAC learning: In **PAC (stands for Probably Approximately Correct) analysis** of a reinforcement learning algorithm, the objective is to bound the sample complexity for obtaining a near-optimal policy. That is, the number of observations a reinforcement learning algorithm has to make before a ϵ-optimal policy is found with probability $1-\delta$. Specific definitions may differ depending on how an ϵ-optimal policy is defined. For example, following Kakade [2003], sample complexity of a PAC algorithm is the number of steps before the value function of any state is within ϵ of the optimal value function. So, an algorithm is PAC efficient if it has low sample complexity with probability $1-\delta$ for any ϵ, δ. The goal is to bound this sample complexity in terms of ϵ, δ, in addition to S, A, and other parameters of the MDP.

- Regret minimization: In **regret analysis**, the objective is to maximize total reward or minimize the difference in reward compared to a benchmark policy, over the steps of the execution of the algorithm. Typically, regret is defined with respect to the performance of best *stationary* policy, which is justified as a near-optimal benchmark if the horizon is large (as discussed earlier there exist an optimal stationary policy for many infinite horizon settings). For example Auer et al. [2009] consider a (weakly) communicating MDP and define regret in T steps as $R(T) = T\rho^* - \sum_{t=1}^{T} r_t$. Here, ρ^* is the optimal gain, which is independent of the starting state and achieved by a stationary policy in this case. And, r_t is the reward obtained by the algorithm at time t. A typical goal is to obtain algorithms that have sublinear in T bounds on regret, in expectation or with high probability.

References

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 89–96. Curran Associates, Inc., 2009. URL `http://papers.nips.cc/paper/3401-near-optimal-regret-bounds-for-reinforcement-learning.pdf`.

Sham M. Kakade. On the sample complexity of reinforcement learning. *PhD thesis, Gatsby Computational Neuroscience Unit, University College London*, 2003.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994. ISBN 0471619779.