Stat RL

Chufan Chen

September 2022

Contents

Chapter 1

Preliminaries

1.1 Markov Decision Process

In reinforcement learning, the interactions between the agent and the environment are often described by a Markov Decision Process (MDP)[2], specified by:

- State space S.
- Action space A.
- Transition function/kernel P: $S \times A \to \Delta(S)$, where $\Delta(S)$ is the space of probability distributions over S (i.e., the probability simplex). $P(s' \mid s, a)$ is the probability of transitioning into state s' upon taking action a in state s.
- Reward function R: $S \times A \to [0, R_{max}]$, where $R_{max} > 0$ is a constant. $R(s, a)$ is the immediate reward associated with taking action a in state s.
- Discount factor $\gamma \in [0, 1)$, which defines a horizon for the problem

1.1.1 Shift of rewards

Why it is sufficient to define reward as nonnegative?
Consider two MDPs $M = (S, A, P, R, \gamma)$ and $M = (S, A, P, R', \gamma)$, which only differ in their reward functions. Moreover, we have for any $s \in S, a \in A$,

$$R(s, a) = R'(s, a) + c$$

, where c is a universal constant that does not depend on s or a. Then these two MDPs in some particular sense are equivalent to each other. For any policy π, let V_M^π denotes its value function in M and $V_{M'}^\pi$, denote its value function in M'. For any $s \in S$,

$$V_M^\pi = V_{M'}^\pi + \frac{c}{1 - \gamma}$$

. The reward shift doesn't change the order of the optimality of the policy and $\forall \pi_1, \pi_2, V_M^{\pi_1} - V_M^{\pi_2} = V_{M'}^{\pi_1} - V_{M'}^{\pi_2}$. In a bounded reward, the constant shift doesn't matter. We can make the assumption that rewards lie in $[0, R_{max}]$ without loss of generality.

1.1.2 Interaction protocol

In a given MDP $M = (S, A, P, R, \gamma)$, the agent interacts with the environment according to the following protocol: the agent starts at some state s_1; at each time step $t = 1, 2, \ldots$, the agent takes an action $a_t \in A$, obtain the immediate reward $r_t = R(s_t, a_t)$, and observes the next state s_{t+1} sampled from $P(s_t, a_t)$, or $s_{t+1} \sim P(s_t, a_t)$. The interaction record

$$\tau = (s_1, a_1, r_1, s_2, \ldots, s_{H+1})$$

is called a trajectory of length H.
In some situations, it is necessary to specify how the initial state s_1 is generated. We consider s_1 sampled from an initial distribution $d_0 \in \Delta(S)$. When d_0 is of importance to the discussion, we include it as part of the MDP definition, and write $M = (S, A, P, R, \gamma, d_0)$.

1.1.3 Finite-horizon MDPs

Without explicit mention, we assume MDP as infinite-horizon discounted for its mathematical convenience. But finite-horizon MDP is more natural in some sense. We cut down the trajectory after H steps, where H is a predefined constant That is, with the same generative process of trajectories, we now consider return to be defined as

$$E[\sum_{h=1}^{H} r_h]$$

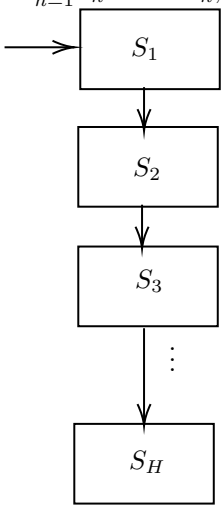
. A finite-horizon MDP is usually specified as $M = (S, A, P, R, H, d_0)$, where H is the episode length (or horizon) and $d_0 \in \Delta(S)$ is the initial state distribution (from which s_1 is drawn). Note that there is no discount factor in finite-horizon MDP. Discount factor is only introduced only for mathematical convenience. It gives you two properties:

1. Stationarity. The optimal policy and value function don't depend on the time-step. It only depends on which state you're in.

2. You can add up an infinite number of rewards that always converge.

Optimal policies in finite-horizon MDPs are generally non-stationary, i.e., you need to look at both the current state and the number of steps remaining to make an optimal decision. In finite-horizon MDPs, you only add up a finite number(H) of rewards, so it doesn't blow up anyway. If we add a constant shift c, $E[\sum_{h=1}^{H}(r_h + c)] = E[\sum_{h=1}^{H} r_h] + cH$. For the value function for every state, consider the following formulation: State space S is layered by a disjoint

H subset, a state in each subset can only appear in a particular time-step: $S = \cup_{h=1}^{H} S_h$. $\forall s \in S_h$, $V_{M'}^{\pi}(s,h) = V_M^{\pi} + c(H-h+1)$.



1.1.4 Indefinite-horizon MDPs

Here is yet another formulation, which is similar to finite-horizon MDPs except that the episode length H can vary: A subset of the state space $S_{term} \subset S$ are considered terminal, and an episode $s_1, a_1, r_1, s_2, a_2, r_2, \ldots$ keeps rolling out until we first visit a terminal state, $s_H \in S_{term}$. In general, the length of the episode, H, is a random variable or a function of policy.The value is still defined as $E[\sum_{h=1}^{H} r_h]$. Examples include the stochastic shortest paths. As an another example, consider consider a navigation task where the goal is to get to the destination state as soon as possible. Let's model it as an indefinite-horizon MDP: reward is -1 per step, and the process terminates whenever we reach the destination. It is clear then the return of a policy is the negative expected total number of steps towards destination. If we add +2 to all rewards, we're not trying to terminate the MDP as soon as possible. You cannot shift reward arbitrary because the length of trajectory depends on how you behave.
Suppose there exists some constant H_0 such that $H \leq H_0$ holds almost surely for an indefinite-horizon MDP. By adding an absorbing state which gives 0 reward and loops in itself. After padding the MDP with the dummy state, you can shift reward both in original MDP states and dummy states.

1.1.5 Non-stationary dynamics

So far all our definitions consider stationary dynamics, that is, the transition function only depends on the state and action, and does not depend on the time step. A finite-horizon MDP with non-stationary dynamics (and reward function) is a generalization: $M = (S, A, \{P_H\}_{h=1}^{H}, H, d_0)$, where $s_1 \sim d_0, s_{h+1} \sim P_h(s_h, a_h)$ and $r_{h+1} = R_h(s_h, a_h)$. That is, the transition rule and reward

function can change as time elapses. Stationary MDP is a special case of non-stationary MDP, $S_1 = S_2 = \ldots = S_H, P_1 = P_2 = \ldots = P_H$. On the other hand, non-stationary MDP is also a special case of stationary MDP, we can augment state space $s' = (s, h)$. Now the state space is H time larger, the benefit is that we can have a single function $P((s', h+1) \mid (s', h), a)$ and not every function for every time-step. By definition, $P((s', h') \mid (s, h), a) = 0$ if $h' - h \neq 1$.

1.1.6 Policy and value

In general, a policy $\pi = (\pi_t)_{t \geq 0}$ is an infinite long sequence where for each $t \geq 0$, $\pi_t : (S \times A)^{t-1} \times S \to M_1(S)$ assigns a probability distribution to histories of length t.(For $\rho > 0$ we use $M_\rho(X)$ to denote the set of nonegative measures μ over X that satisfy $\mu(X) = \rho$.) A (deterministic and stationary) policy $\pi : S \to A$ specifies a decision-making strategy in which the agent chooses actions adaptively based on the current state, i.e., $a_t = \pi(s_t)$. More generally, the agent may also choose actions according to a stochastic policy $\pi : S \to \Delta(A)$, and with a slight abuse of notation we write $a_t \sim \pi(s_t)$. A deterministic policy is its special case when $\pi(s)$ is a point mass for all $s \in S$.
The goal of the agent is to choose a policy π to maximize the expected discounted sum of rewards, or value:

$$E[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid \pi, s_1]. \tag{1.1}$$

The expectation is with respect to the randomness of the trajectory, that is, the randomness in state transitions and the stochasticity of π. Notice that, since r_t is nonnegative and upper bounded by R_{max}, we have

$$0 \leq \sum_{t=1}^{\infty} \gamma^{t-1} r_t \leq \sum_{t=1}^{\infty} \gamma^{t-1} R_{max} = \frac{R_{max}}{1-\gamma}. \tag{1.2}$$

Hence, the discounted sum of rewards (or the discounted return) along any actual trajectory is always bounded in the range $[0, \frac{R_{max}}{1-\gamma}]$, and so is its expectation of any form. This fact will be important when we later analyze the error propagation of planning and learning algorithms.
Note that for a fixed policy, its value may differ for a different choice of s_1, and we define the value function $V_M^\pi : S \to \mathbb{R}$ as

$$V_M^\pi(s) = E[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid \pi, s_1 = s],$$

which is the value obtained by following policy π starting at state s. Similarly, we define the action-value(or Q-value) function $Q_M^\pi : S \times A \to \mathbb{R}$ as

$$Q_M^\pi(s, a) = E[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid \pi, s_1 = s, a_1 = a].$$

Henceforth, the dependence of any notation on M will be made implicit whenever it is clear from the context.

1.1.7 Bellman equation for policy evaluation

Based on the principles of dynamic programming, V^π and Q^π can be computed using the following bellman equation for policy evaluation: $\forall s \in S, a \in A$

$$\begin{aligned} V^\pi(s) &= E[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid s_1 = s, \pi] \\ &= Q(s, \pi(s)) \\ &= R(s, \pi(s)) + \gamma E_{s' \sim P(s,a)}[V^\pi(s')] \\ &= R(s, \pi(s)) + \gamma \sum_{s' \in S} P(s' \mid s, \pi(s)) V^\pi(s') \end{aligned} \tag{1.3}$$

$$\begin{aligned} Q^\pi(s, a) &= R(s, a) + \gamma E_{s' \sim P(s,a)}[V^\pi(s')] \\ &= R(s, a) + \gamma E_{s' \sim P(s,a)}[Q^\pi(s', \pi(s'))] \end{aligned} \tag{1.4}$$

In $Q(s, \pi(s))$ and $R(s, \pi(s))$ we treat π as a deterministic policy for brevity, and for stochastic policies this shorthand should be interpreted as $E_{a \sim \pi(s)}[Q^\pi(s, a)]$ and $E_{a \sim \pi(s)}[R^\pi(s, a)]$.
If we assume S and A are finite, upon fixing an arbitrary order of states and actions, we can rewrite 1.3 and 1.4 in matrix form and derive an analytical solution for V^π and Q^π using linear algebra as below. Define:

- V^π as the $\mid S \mid \times 1$ vector $[V^\pi(s)]_{s \in S}$
- R^π as the reward vector for policy π with dimension $\mid S \mid \times 1$, whose s-th entry is
$$[R^\pi]_s = E_{a \sim \pi(s)}[R(s, a)]$$
. For deterministic policy π,
$$[R^\pi]_s = R(s, \pi(s))$$
.
- P^π as the matrix $[P(s' \mid s, \pi(s))]_{s \in S, s' \in S}$, whose (s, s')-th entry is
$$[P^\pi]_{s,s'} = E_{a \sim \pi(s)}[P(s' \mid s, a)]$$
. Similarly, for deterministic policy π,
$$[P^\pi]_{s,s'} = P(s' \mid s, \pi(s))$$
.In fact, this matrix describes a Markov chain induced by MDP M and policy π. Its s-th row is the distribution over next-states upon taking actions according to π at state s , which we also write as $[P(s, \pi)]^\top$.

Then from 1.3 we have

$$\begin{aligned} \forall s, [V^\pi]_s &= [R^\pi]_s + \gamma \langle P(s, \pi), V^\pi \rangle \\ V^\pi &= R^\pi + \gamma P^\pi V^\pi \end{aligned} \tag{1.5}$$

Lemma 1.1.1 (Bellman consistency)**.** *For stationary polices, we have*

$$V^\pi = Q^\pi(s, \pi(s)) = \mathbb{E}_{a\sim\pi_{a|s}}$$

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

In matrix forms:

$$V^\pi = R^\pi + \gamma P^\pi V^\pi \quad P^\pi \in \mathbb{R}^{S\times S}$$

$$Q^\pi = R + \gamma P V^\pi$$

$$Q^\pi = R + \gamma P^\pi Q^\pi \quad P^\pi \in \mathbb{R}^{SA\times SA}$$

where $R \in \mathbb{R}^{SA}, R^\pi \in \mathbb{R}^S$.

$$\begin{aligned} V^\pi &= R^\pi + \gamma P^\pi V^\pi \\ (I - \gamma P^\pi)V^\pi &= R^\pi \\ V^\pi &= (I - \gamma P^\pi)^{-1} R^\pi \end{aligned}$$

Now we notice that matrix$(I_{|S|} - \gamma P^\pi)$ is always invertible/full rank for any $\gamma < 1$. Ways to prove a matrix A is invertible:

1. The null space/kernel of A is trivial. That is $Ax = 0 \iff x = 0$.
2. 0 is not an eigenvalue.
3. The determinant of A is nonzero.
4. If at any point of the Gauss-Jordan process on A you can get it into a reduced row echelon form. Likewise, A is not invertible if you get a row or column of all zeros at some point while using Gauss-Jordan.

Here we use the first method:

$$\begin{aligned} \|(I - \gamma P^\pi)x\|_\infty &\geq \|Ix\|_\infty - \gamma\|P^\pi x\|_\infty \text{(triangular inequality for norms)} \\ &\geq \|Ix\|_\infty - \gamma\|x\|_\infty \text{(each element of } P^\pi x \text{ is a convex average of x)} \\ &= (1-\gamma)\|x\|_\infty \geq 0 \end{aligned} \tag{1.6}$$

So we can conclude that

$$V^\pi = (I - \gamma P^\pi)^{-1} R^\pi. \tag{1.7}$$

$$P^\pi x = \begin{bmatrix} \langle P^\pi[1,:], x\rangle \\ \vdots \\ \langle P^\pi[n,:], x\rangle \end{bmatrix}$$

By Holder's inequality, $\langle P^\pi[n,:], x\rangle \leq \|P^\pi[m,:]\|_1 \|x\|_\infty \leq \|x\|_\infty$.
Why does the infinity norm work here?
P is a row-stochastic matrix. Hölder's inequality tells us that $l1$-norm and l-infinity norm are dual pairs.

1.1.8 State occupancy

$$(I - \gamma P^{\pi})^{-1}$$

is a matrix, where each row (indexed by s) is the discounted state occupancy d_s^{π}, whose s'-th entry is

$$d_s^{\pi}(s') = E[\sum_{t=1}^{\infty} \gamma^{t-1} \mathbb{1}[s_t = s'] \mid s_1 = s, \pi]$$

. It is similar to the notion of occupancy distribution/stationary distribution in the ergodic Markov chain.

1. Each row is like a distribution vector except that the entries sum up to $\frac{1}{1-\gamma}$. Let $\eta_s^{\pi} = (1-\gamma)d_s^{\pi}$ denote the normalized vector.
2. $V^{\pi}(s)$ is the dot product between d_s^{π} and reward vector.
3. Can also be interpreted as the value function of the indicator reward function.

Similaryly, we can define state-action occupancy $d^{\pi}(S_t = s, A_t = a)$.

1.1.9 Optimality

The standard goal in an MDP is to identify a policy that maximizes this value in every state. A policy achieving this is known as an optimal policy. Whether an optimal policy exists at all is not clear at this stage. In any case, if it exist, an optimal policy must satisfy $V^{\pi} = V^*$ where $V^* : S \to \mathbb{R}$ is defined by

$$V^*(s) = \sup_{\pi \in \Pi} V^{\pi}(s), s \in S.$$

Let $\epsilon > 0$. A policy π is said to be ϵ-optimal if

$$V^{\pi} \geq V^* - \epsilon \vec{1}.$$

Finding an ϵ-optimal policy with a positive ϵ should intuitively be easier than finding an optimal policy.

1. Understand why such π^* exists?
2. What does π^* looks like?

Proposition 1.1.1.1. *It suffices to consider stationary(memoryless) policies.*

Proof. It suffices to show that:
For any π, $\exists \pi'$ such that π' is stationary and both π and π' have the same value:

$V^{\pi}(\mu) = V^{\pi'}(\mu)$.
By definition:

$$\begin{aligned} V^{\pi}(\mu) &= E_s^{\pi}[\sum_{t=1}^{\infty} \gamma^{t-1} r(S_t, A_t)] \\ &= \sum_{t=1}^{\infty} \gamma^{t-1} E_{\mu}^{\pi}[r(S_t, A_t)] \\ &= \sum_{t=1}^{\infty} \gamma^{t-1} \sum_{s,a} P_{\mu}^{\pi}(S_t = s, A_t = a) r(s, a) \\ &= \sum_{s,a} [\sum_{t=1}^{\infty} \gamma^{t-1} P_{\mu}^{\pi}(S_t = s, A_t = a)] r(s, a) \\ &= \sum_{s,a} \nu_{\mu}^{\pi}(s, a) r(s, a) \end{aligned}$$

Here, $\nu_{\mu}^{\pi}(s, a)$ is termed as the occupancy measure

$$\nu_{\mu}^{\pi}(s, a) = \sum_{t=1}^{\infty} \gamma^{t-1} d^{\pi}(S_t = s, A_t = a),$$

$$\nu_{\mu}^{\pi}(s) = \sum_{t=1}^{\infty} \gamma^{t-1} d^{\pi}(S_t = s),$$

where $d^{\pi}(S_t = s)$ is marginal density function under policy π at time t observe state s and $d^{\pi}(S_t = s, A_t = a)$ is marginal distribution function under policy π at time t observe state-action pair (s,a). And the value function can alternatively be written in terms of occupancy measure as: $V_{\mu}^{\pi} = \langle \nu_{\mu}^{\pi}(s, a), r \rangle$. From above, we conclude that it suffices to prove that the occupancy measures with both policies are equivalent: $\nu_{\mu}^{\pi} = \nu_{\mu}^{\pi'}$. Consider π' for any fixed π. Define:

$$\pi'(a \mid s) = \begin{cases} \frac{\nu_{\mu}^{\pi}(s,a)}{\nu_{\mu}^{\pi}(s)} & \text{if } \nu_{\mu}^{\pi}(s) > 0 \\ \pi_0(a) & \text{otherwise} \end{cases} \tag{1.8}$$

where $\nu_{\mu}^{\pi}(s) = \sum_a \nu_{\mu}^{\pi}(s, a)$. It is clear that π' is stationary, $\pi' : S \to \Delta(A)$. By law of total probability and 1.8:

$$P^{\pi'}(s' \mid s) = \sum_a \pi'(a \mid s) P(s' \mid s, a) = \sum_a \frac{\nu_{\mu}^{\pi}(s, a)}{\nu_{\mu}^{\pi}(s)} P(s' \mid s, a) \tag{1.9}$$

From the definition of occupancy measure, we had:

$$\begin{aligned}
\nu_\mu^\pi(s) &= \sum_{t=1}^{\infty} \gamma^{t-1} P_\mu^\pi(S_t = s) \\
&= \mu(s) + \gamma \sum_{t=2}^{\infty} \gamma^{t-2} P_\mu^\pi P_\mu^\pi(S_t = s) \\
&= \mu(s) + \gamma \sum_{\tilde{t}=1}^{\infty} \gamma^{\tilde{t}-1} P_\mu^\pi P_\mu^\pi(S_{\tilde{t}+1} = s) \quad [\tilde{t} = t-1] \\
&= \mu(s) + \gamma \sum_{t=1}^{\infty} \gamma^{t-1} \sum_{s',a} P_\mu^\pi(S_t = s', A_t = a) P(s \mid s', a) \\
&= \mu(s) + \gamma \sum_{s',a} [\sum_{t=1}^{\infty} \gamma^{t-1} P_\mu^\pi(S_t = s', A_t = a)] P(s \mid s', a) \\
&= \mu(s) + \gamma \sum_{s',a} \nu_\mu^\pi(s', a) P(s \mid s', a)
\end{aligned}$$

By multiplying and dividing $\nu_\mu^\pi(s')$ and using 1.9, we get:

$$\nu_\mu^\pi(s) = \mu(s) + \gamma \sum_{s'} \nu_\mu^\pi(s') P^{\pi'}(s \mid s') \Rightarrow \nu_\mu^\pi = (I - \gamma P^{\pi'})^{-1} \mu$$

. Similarly,

$$\nu_\mu^{\pi'}(s) = \mu(s) + \gamma \sum_{s'} \nu_\mu^\pi(s') P^{\pi'}(s \mid s') \Rightarrow \nu_\mu^{\pi'} = (I - \gamma P^{\pi'})^{-1} \mu$$

The occupancy measures and hence the value function corresponding to any general policy and the constructed stationary policy are shown to be equal. □

Corollary 1.1.1.1. *There exists an optimal policy that is stationary.*

Proposition 1.1.1.2. *For infinite-horizon discounted MDPs, a stationary and deterministic policy always exists that is optimal for all starting states simultaneously.*

Let π^* denote this optimal policy, and $V^* := V^{\pi^*}$. Bellman optimality equation:

$$V^*(s) = \max_{a \in A} Q^*(s, a) = \max_{a \in A} (R(s, a) + \gamma E_{s' \sim P(s,a)}[V^*(s')]) \tag{1.10}$$

. Different from the bellman equation:

1. no reference to any policy π.
2. the max operator which makes this a nonlinear set of equations

Given the value function V, a greedy policy π_V can be defined by selecting for each state the action that maximizes the state's value, i.e.,

$$\pi_V(s) = arg\max_{a \in A}[R(s,a) + \gamma \sum_{s' \in S} P_{ss'}(a)V(s')],$$

where ties for the maximum action are broken arbitrarily. Evaluating a greedy policy π_V yields a new value function V_{π_V}, which we abbreviate as V_V. Value function V gives rise to greedy policy π_V which, when evaluated, yields V_V. In general, $V \neq V_V$. Equality occurs if and only if $V = V^*$, in which case any greedy policy will be optimal.
If we know V^*, how to get π^*?
Compute the RHS in the bracket of the bellman optimality equation for every single action and take the argmax. Sometimes it's not available in reinforcement learning, because we don't the transition dynamics in the learning setting.
In reinforcement learning, it's easier to work with Q-values: $Q^*(s,a)$, as $\pi^*(s) = arg\max_{a \in A} Q^*(s,a)$.

$$Q^*(s,a) = R(s,a) + \gamma E_{s' \sim P(s,a)}[\max_{a' \in A} Q^*(s',a')] \tag{1.11}$$

Proposition 1.1.1.3. *There is a deterministic, stationary and optimal policy and it is given by* $\pi^*(s) = arg\max_a Q^*(s,a)$.

Proof. π^* is stationary.

$$\begin{aligned} V^{\pi'}(s) \leq V^*(s) = V^{\pi^*}(s) &= \mathbb{E}_{a \sim \pi^*(a|s)}[Q^{\pi^*}(s,a)] \\ &\leq \max_a Q^{\pi^*}(s,a) \\ &= \max_a Q^*(s,a) \\ &= Q(s,\pi'(s)) = V^{\pi'}(s) \quad \text{define } \pi'(s) = arg\max_a Q^*(s,a) \end{aligned}$$

1. Check π' is stationary
2. π' is deterministic

□

We often say the policy is a greedy policy of the particular Q-function by the way of inducing a policy from a state-action function/Q-function. $\pi^*(s)$ is the greedy policy with respect to Q^*. We use shorthand π_Q to denote the procedure of turning a Q-value function into its greedy policy, and the above equation can be written as

$$\pi^* = \pi_{Q^*} := (s \mapsto arg\max_{a \in A} Q^*(s,a))$$

.

$$V^* = V^{\pi^*}, Q^* = Q^{\pi^*}$$

V^* and Q^* are uniquely defined, but the π^* is not necessarily unique. For example, consider an MDP where the immediate rewards are zero everywhere,

then any policy is optimal. Bellman operators:$\mathcal{T} : \mathbb{R}^{S\times A} \to \mathbb{R}^{S\times A}$, $\forall f \in \mathbb{R}^{S\times A}$, $(\mathcal{T}f)(s,a) = R(s,a) + \gamma E_{s'\sim P(s,a)}[V_f(s')]$, where $V_f(s') := \max_{a'} f(s',a')$. This allows us to rewrite Bellman Optimality Equation 1.11 in the following concise form, which implies that Q^* is the fixed point of the operator $\mathcal{T}$:

$$Q^* = \mathcal{T}Q^*$$

. Similarly define $\mathcal{T}^\pi : \mathbb{R}^{S\times A} \to \mathbb{R}^{S\times A}$, $(\mathcal{T}^\pi f)(s,a) = R(s,a)+\gamma E_{s'\sim P(s,a)}[f(s',\pi(s'))]$. We can rewrite the Bellman Equation for Policy Evaluation 1.4 to

$$Q^\pi = \mathcal{T}^\pi Q^\pi$$

. With abuse of notation, we can write bellman equation for value function as $V^* = \mathcal{T}V^*$ and $V^\pi = \mathcal{T}^\pi V^\pi$. Strictly speaking, the $\mathcal{T}$ and $\mathcal{T}^\pi$ here are different from the equation above.

1.1.10 Evaluation error to decision loss

There are K items, $1, 2, \ldots, K$. The i-th item has value $v_i \in \mathbb{R}$. Let $v^* := \max_{i\in[K]} v_i$, where $[K] := \{1, 2, \ldots, K\}$, and $i^* := arg\max_{i\in[K]} v_i$. An agent chooses the j-th item, where $j := arg\max_{i\in[K]} u_i$, and $\{u_i\}_{i=1}^K$ are K real numbers. Let

$$\epsilon := \max_{i\in[K]} |u_i - v_i|.$$

Upper-bound $v^* - v_j$ as a function of ϵ.

Proof.

$$\begin{aligned} v^* - v_j = |v^* - v_j| &= |v^* - u_{i^*} + u_{i^*} - v_j| \\ &\le |v^* - u_{i^*}| + |u_{i^*} - v_j| \\ &\le \epsilon + |u_{i^*} - v_j| \end{aligned}$$

□

If we further have $\forall i, u_i \le v_i$, the upper bound is

$$v^* - v_j = v^* - u_{i^*} - v_j + u_{i^*} - v_j \le v^* - u_{i^*} - v_j + u_j - v_j \le 2\epsilon$$

. It's the same for $\forall i, u_i \ge v_i$. The upper bound of $|u_j - v^*|$ is $u_j - u_{i^*} + \epsilon$.

1.2 Planning in MDPs

Planning means we want to compute V^π, Q^π, V^*, Q^* given the MDP model. Value iteration and policy iteration are two basic algorithms for planning.

1.2.1 Value Iteration

In value iteration, we are interested in computing Q^* and it satisfies the bellman optimality equation $Q^* = \mathcal{T}Q^*$. Recall that $\mathcal{T}$ is the Bellman optimality operator. In general, the equation is a fix point equation, and $\mathcal{T}$ is the fix point operator. This kind of problem is studied extensively in numerical analysis. One of the most basic ideas you can try here is the power method which starts with an arbitrary function and keeps applying the operator. If the operator satisfies some property, it's getting closer and closer to the real function. We'll show that $\mathcal{T}$ satisfies the contraction property in the MDP setting.
The algorithm: define $Q^{*,0} := \vec{0} \in \mathbb{R}^{S\times A}$.

$$Q^{*,h} := \mathcal{T}Q^{*,h-1}$$

, stop at large $h = H$. There are two questions:

- How large is $\|Q^* - Q^{*,H}\|$?
- Can I extract a good policy from $Q^{*,H}$? How good is this policy $\pi_{Q^{*,H}}$?

$$\pi_h := \pi_{Q^{*,h}}, Q^{*,h} \neq Q^{\pi_h}$$

. Let's answer the second question first. I have an arbitrary $f \in \mathbb{R}^{S\times A}$. I'll act using π_f in M. Assume that I know that $\|f - Q^*\|$ is small. Can I bound $V^* - V^{\pi_f}$?

Lemma 1.2.1 ([3])**.**

$$\|V^* - V^{\pi_f}\|_\infty \leq \frac{2\|f - Q^*\|_\infty}{1-\gamma}$$

Proof. $\|V^* - V^{\pi_f}\|_\infty$ is always non-negative by definition. $\forall s$,

$$\begin{aligned}
V^*(s) - V^{\pi}(s) &= Q^*(s,\pi^*(s)) - Q^{\pi_f}(s,\pi_f(s)) \\
&= Q^*(s,\pi^*(s)) - Q^*(s,\pi_f(s)) + Q^*(s,\pi_f(s)) - Q^{\pi_f}(s,\pi_f(s)) \\
&\leq Q^*(s,\pi^*(s)) - f(s,\pi^*(s)) + f(s,\pi_f(s)) - Q^*(s,\pi_f(s)) \\
&+ R(s,\pi_f(s)) + \gamma E_{s'\sim P(s,\pi_f)}[V^*(s')] - R(s,\pi_f(s)) - \gamma E_{s'\sim P(s,\pi_f)}[V^{\pi_f}(s')] \\
&\leq Q*(s,\pi^*(s)) - f(s,\pi^*(s)) + f(s,\pi_f(s)) - Q^*(s,\pi_f(s)) \\
&+ \gamma E_{s'\sim P(s,\pi_f)}[V^*(s') - V^{\pi_f}(s')] \\
&\leq Q^*(s,\pi^*(s)) - f(s,\pi^*(s)) + f(s,\pi_f(s)) - Q^*(s,\pi_f(s)) + \gamma\|V^* - V^{\pi_f}\|_\infty \\
&\leq 2\|f - Q^*\|_\infty + \gamma\|V^* - V^{\pi_f}\|_\infty
\end{aligned}$$

So,

$$\|V^* - V^{\pi_f}\|_\infty \leq 2\|f - Q^*\|_\infty + \gamma\|V^* - V^{\pi_f}\|_\infty$$

$f(s,\pi_f(s)) \geq f(s,\pi^*(s'))$ holds because by definiton: $\pi_f = arg\max_{a\in A} f(s,a)$, $f(s,\pi_f(s)) = \max_{a\in A} f(s,a) \geq f(s,a'), \forall a' \in A$. □

For finite-horizon

$$\|V^* - V^{\pi_f}\|_\infty \leq 2H\|f - Q^*\|_\infty$$

$$\frac{1}{1-\gamma} \sim H$$

. Then it remains to show that by performing value iteration for a large number of iterations, we'll get a good approximation of Q^* under the infinity norm. Goal: bound $\|Q^{*,H} - Q^*\|_\infty$.

Lemma 1.2.2. *$\forall f, f'$: $\|\mathcal{T}f - \mathcal{T}f'\|_\infty \leq \gamma\|f - f'\|_\infty$. Formally We say that the $\mathcal{T}$ operator is a γ contraction under the L-infinity norm.*

The lemma is useful because $\forall h \geq 1$,

$$\begin{aligned}\|Q^{*,H} - Q^*\|_\infty &= \|\mathcal{T}Q^{*,H-1} - \mathcal{T}Q^*\|_\infty \\ &\leq \gamma\|Q^{*,H-1} - Q^*\|_\infty \\ &\leq \gamma^H\|Q^{*,0} - Q^*\|_\infty \\ &\leq \gamma^H \frac{R_{max}}{1-\gamma}\end{aligned}$$

How about using a small γ to compute a value function and policy as a warmup/starting point?
Then all it remains is to prove the contraction property. As a side comment, given that we have the bound we can easily backup the H to achieve a certain desired bound. For example, I want $\|Q^{*,H} - Q^*\|_\infty \leq \epsilon$. How large H needs to be? The answer is $H \geq \frac{1}{1-\gamma}\log\frac{R_{max}}{\epsilon(1-\gamma)}$. When we prove some infinity norm is upper bound by some infinity norm else. For LHS things, what you particularly do is consider the function you evaluate at any state/state-action pair and bound that uses the RHS at every single state/state-action pair, then you can replace the LHS with infinity norm.

Proof. $\forall(s,a)$,

$$\begin{aligned}&|(\mathcal{T}f)(s,a) - (\mathcal{T}f')(s,a)| \\ &\qquad = |R(s,a) + \gamma E_{s'\sim P(s,a)}[V_f(s')] - R(s,a) - \gamma E_{s'\sim P(s,a)}[V_{f'}(s')]| \\ &\qquad = \gamma|E_{s'\sim P(s,a)}[V_f(s') - V_{f'}(s')]| \\ &\qquad \leq \gamma\|V_f - V_{f'}\|_\infty\end{aligned}$$

Then we want to show that $\gamma\|V_f - V_{f'}\|_\infty \leq \|f - f'\|_\infty$. Note that V_f and $V_{f'}$ are state value function($\mathbb{R}^S$) whereas f and f' are state-action value function($\mathbb{R}^{S\times A}$). It suffices to show: $\forall s\ |V_f(s) - V_{f'}(s)| \leq \max_a |f(s,a) - f'(s,a)|$.

$$\begin{aligned}&|V_f(s) - V_{f'}(s)| \leq \max_a |f(s,a) - f'(s,a)| \\ \Leftrightarrow\ &|\max_a f(s,a) - \max_a f'(s,a)| \leq \max_a |f(s,a) - f'(s,a)|\end{aligned}$$

w.l.o.g we can assume $V_f(s) \geq V_{f'}(s)$ and define $a^* = arg\max_a f(s,a)$ (If we assume $V_f(s) \leq V_{f'}(s)$ now we need to define $a^* = arg\max_a f'(s,a)$). So, $f(s,a^*) = V_f(s) = \max_a f(s,a)$.

$$\begin{aligned}\max_a f(s,a) - \max_a f'(s,a) &= f(s,a^*) - \max_a f'(s,a) \\ &\leq f(s,a^*) - f'(s,a^*) \\ &= |f(s,a^*) - f'(s,a^*)| \\ &\leq \|f - f'\|_\infty\end{aligned}$$

□

An alternative way of bounding $\|Q^{*,H} - Q^*\|_\infty$.We borrow the concept from the finite-horizon MDP and bring a new informal definition of value function(H truncated value function)

$$V^{\pi,H}(s) := E[\sum_{t=1}^{H} \gamma^{t-1} r_t \mid \pi, s_1 = s]$$

. We only care about the maximum value obtained from the state:

$$V^{*,H}(s) := \max_\pi V^{\pi,H}(s)$$

. Similarly we can define $Q^{*,H}(s,a)$.
Claim: $Q^{*,H}$ as the output of VI is the optimal Q-function for the H-step truncated objective.

$$Q^{*,0} = \vec{0}.Q^{*,1} = \mathcal{T}Q^{*,0} = R.Q^{*,2} = R(s,a) + \gamma E_{s' \sim P(s,a)}[max_{a'} R(s',a')]$$

$$V^{*,H})(s) = \max_a Q^{*,H}(s,a)$$

Note that π^* is optimal wrt $\sum_{t=1}^{\infty} \gamma^{t-1} r_t$.$\pi^{*,H}$ is optimal wrt $\sum_{t=1}^{H} \gamma^{t-1} r_t$. So,

$$Q^{\pi^*,H} \leq Q^{*,H}$$

$$\begin{aligned}0 \leq Q^* - Q^{*,H} &\leq Q^* - Q^{\pi^*,H} \\ &= Q^{\pi^*} - Q^{\pi^*,H}\end{aligned}$$

$\forall s, a,$

$$
\begin{aligned}
& Q^{\pi^*}(s,a) - Q^{\pi^*,H}(s,a) \\
&= E[(\sum_{t=1}^{\infty} \gamma^{t-1} r_t) - (\sum_{t=1}^{H} \gamma^{t-1} r_t) \mid s_1 = s, a_1 = a, \pi^*] \\
&= E[\sum_{t=H+1}^{\infty} \gamma^{t-1} r_t \mid s_1 = s, a_1 = a, \pi^*] \\
&\leq \gamma^H (\sum_{t=1}^{\infty} \gamma^{t-1} R_{max}) \\
&= \frac{\gamma^H R_{max}}{1-\gamma} = R_{max} \frac{(1-(1-\gamma))^H}{1-\gamma} \\
&\leq R_{max} \frac{e^{-(1-\gamma)^k}}{1-\gamma}
\end{aligned}
$$

The last inequality uses

$$\lim_{n\to\infty} (1 - \frac{1}{n})^n = e^{-1} \Rightarrow (1 - \frac{1}{n})^n \leq e^{-1} \quad \forall n$$

We can compute iteration complexity from the convergence bound. Set $\epsilon = \frac{e^{-(1-\gamma)^k}}{1-\gamma}$, the solve this equation to get $k = \frac{\log \epsilon(1-\gamma)}{-(1-\gamma)}$.Convergence of the Q function implies the convergence of the value of the induced policy.
If $Q^{*,0}$ is not a zero vector. We modify the definition as

$$V^{\pi,H}(s) := E[\sum_{t=1}^{H} \gamma^{t-1} r_t + \gamma^H Q^{*,0}(s_H, a_H) \mid \pi, s_1 = s]$$

. In other words, $Q^{*,0} = \vec{0}$ is a special case where $Q^{*,0}(s_H, a_H) = 0$.

1.2.2 Policy Iteration

Initial π_0 arbitrarily, and repeat the following iterative procedure: for $k = 1, 2, \ldots$, $\pi_k \leftarrow \pi_{Q^{\pi_{k-1}}}$.

1. Compute the $Q^{\pi_{k-1}}$. Policy Evaluation Step
2. Take greedy policy. Policy Improvement Step

Assume we can solve Q^π for given π. How?

$$Q^\pi = \mathcal{T}^\pi Q^\pi.$$

$$(\mathcal{T}^\pi f)(s,a) = R(s,a) + \gamma E_{s' \sim P(s,a)}[f(s', \pi)].$$

The above is a linear equation that can solve by matrix inverse. Alternatively, start with arbitrarily f_0, $f_i \leftarrow \mathcal{T}^\pi f_{i-1}$(works because $\mathcal{T}^\pi$ is γ-contraction under l_∞).

Theorem 1.2.3 (Policy improvement theorem)**.** $V^{\pi_k} \geq V^{\pi_{k-1}}$. *Furthermore, unless* $\pi_k = \pi^*$ *, improvement in at least 1 state is non-zero.[2]*

Therefore, the termination criterion for policy iteration is $Q^{\pi_k} = Q^{\pi_{k-1}}$. Since we are only searching over stationary and deterministic policies, a new policy that is different from all previous ones is found in every iteration.

Corollary 1.2.3.1. *PI terminates in at most* $|A|^{|S|}$ *iterations.*

Proof. Claim: "Monotonicity of $\mathcal{T}$": $\forall f \leq f', \mathcal{T}f \leq \mathcal{T}f'$. By definition:

$$(\mathcal{T}f)(s,a) = R(s,a) + \gamma E_{s' \sim P(s,a)}[\max_{a'} f(s',a')],$$

$$v_f := \max_{a'} f(s',a').$$

This claim holds because $v_f \leq v_{f'}$.

$$Q^{\pi_k} = \mathcal{T}^{\pi_k} Q^{\pi_k} \leq \mathcal{T} Q^{\pi_k} = \mathcal{T}^{\pi_{k+1}} Q^{\pi_k}$$

1. $\forall f, \pi, \mathcal{T}^{\pi} f \leq \mathcal{T} f$.
2. $\pi_{k+1} = \pi_{Q^{\pi_k}}$.
3. $\forall f, \mathcal{T}f = \mathcal{T}^{\pi_f} f.(\max_{a'} f(s',a') = f(s',\pi_f))$

$$\begin{aligned} Q^{\pi_k} &\leq \mathcal{T}^{\pi_{k+1}} Q^{\pi_k} \\ &\leq \mathcal{T}^{\pi_{k+1}} (\mathcal{T}^{\pi_{k+1}} Q^{\pi_k}) \text{(policy evaluation operator also has monotonicity property)} \\ &\leq (\mathcal{T}^{\pi_{k+1}})^{\infty} Q^{\pi_k} \\ &= Q^{\pi_{k+1}} (Q^{\pi_{k+1}} \text{ is the fix point of } \mathcal{T}^{\pi_{k+1}}) \end{aligned}$$

□

Alternative Proof.

Definition 1.2.3.1 (Advantage)**.** $A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s)$. $A^{\pi}(s,\pi) = 0$. *It measures the deviation of choosing action* a *and then following the policy* π *from following policy* π*. The advantage of policy* π' *over policy* π *is defined as* $A^{\pi}(s,\pi') := A^{\pi}(s,\pi'(s))$.

Since policy iteration always takes the greedy policy of the current policy's Q-value function, by definition the advantage of the new policy over the old one is non-negative. The next result shows that the value difference between two policies can be expressed using the advantage function. The policy improvement theorem immediately follows, since $V^{\pi_k}(s) - V^{\pi_{k-1}}(s)$ can be decomposed into the sum of nonnegative terms.

Lemma 1.2.4 (Performance-difference lemma[1])**.** $\forall \pi', \pi$*, and any state* $s \in S$*,*

$$V^{\pi'}(s) - V^{\pi}(s) = \frac{1}{1-\gamma} E_{s' \sim d^{\pi',s}}[A^{\pi}(s',\pi')],$$

where $d^{\pi',s}$ *is the normalized occupancy of* π' *with* s *as initial state.*
$d^{\pi} = (1-\gamma)\sum_{t=1}^{\infty} \gamma^{t-1} d_t^{\pi}$*,* d_t^{π} *is the distribution of* s_t *under* π*, starting from* d_0*.*

$$V^{\pi_{k+1}}(s) \geq V^{\pi_k}(s) \Leftrightarrow V^{\pi_{k+1}}(s) - V^{\pi_k}(s) \geq 0$$
$$\Leftrightarrow \frac{1}{1-\gamma} E_{s' \sim d^{\pi_{k+1}, s}}[A^{\pi_k}(s', \pi_{k+1}(s'))] \geq 0$$

$$A^{\pi_k}(s', \pi_{k+1}(s')) = Q^{\pi_k}(s', \pi_{k+1}(s')) - V^{\pi_k}(s')$$
$$= \max_{a'} Q^{\pi_k}(s', a') - Q^{\pi_k}(s', \pi_k) \geq 0$$

□

Strictness. Towards contradiction: $V^{\pi_{k+1}} = V^{\pi_k}$. $\forall s'$,

$$A^{\pi_k}(s', \pi_{k+1}) = 0$$

.

$$Q^{\pi_k}(s', \pi_k) = \max_{a'} Q^{\pi_k}(s', a')$$

$$\forall s, V^{\pi_{k+1}}(s) = V^{\pi_k}(s) \xrightarrow{\text{we want to prove}} \pi_k = \pi^*$$

$$\forall s, E_{s' \sim d^{\pi, s}}[A^{\pi_k}(s', \pi_{k+1}])] = 0$$

Claim:

$$A^{\pi_k}(s', \pi_{k+1}) = 0. \forall s'$$

. What if $\exists s$, $A^{\pi_k}(s, \pi_{k+1}) > 0$. $\Rightarrow V^{\pi_{k+1}}(s) - V^{\pi_k}(s) = \frac{1}{1-\gamma} E_{s' \sim d^{\pi, s}}[A^{\pi_k}(s', \pi_{k+1})]$. Remind that $d^{\pi, s} = (1-\gamma) \sum_{t=1}^{\infty} \gamma^{t-1} d_t^{\pi, s}$.

$$V^{\pi_{k+1}}(s) - V^{\pi_k}(s) = \frac{1}{1-\gamma} E_{s' \sim d^{\pi, s}}[A^{\pi_k}(s', \pi_{k+1})]$$
$$\geq E_{s' \sim d_1^{\pi, s}}[A^{\pi_k}(s', \pi_{k+1})]$$
$$= A^{\pi_k}(s, \pi_{k+1}) > 0$$

$$\max_{a'} Q^{\pi_k}(s', a') = Q^{\pi_k}(s', \pi_{k+1}) = Q^{\pi_k}(s', \pi_k), \forall s'.$$

$$Q^{\pi_k} = \mathcal{T}^{\pi_k} Q^{\pi_k} = \mathcal{T} Q^{\pi_k}. Q^{\pi_k} = Q^*.$$

□

Bibliography

[1] Sham Kakade, Sham Kakade, and John Langford. "Approximately Optimal Approximate Reinforcement Learning". In: *IN PROC. 19TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING* (2002), pp. 267–274. URL: http://130.203.136.95/viewdoc/summary?doi=10.1.1.7.7601.

[2] Martin L. Puterman. "Markov decision processes : discrete stochastic dynamic programming". In: (2005), p. 649. URL: https://www.wiley.com/en-us/Markov+Decision+Processes%3A+Discrete+Stochastic+Dynamic+Programming-p-9780471727828.

[3] Satinder P. Singh and Richard C. Yee. "An upper bound on the loss from approximate optimal-value functions". In: *Machine Learning 1994 16:3* 16.3 (Sept. 1994), pp. 227–233. ISSN: 1573-0565. DOI: 10.1007/BF00993308. URL: https://link.springer.com/article/10.1007/BF00993308.