Notes on Reinforcement Learning

Suei-Wen Chen

Sep 13, 2023

Table of Contents

List of Symbols

G	Return
G_t	Future return from time t
R	Reward
T	Time of terminatin of an episode
Π	The set of all policies
Π_{det}	The set of all deterministic policies
Π_{stat}	The set of all stationary policies
$\mathcal{A}(x)$	Admissible action at state x
$\mathcal{A}$	Action space
$\mathcal{P}$	State transition probability kernel
$\mathcal{P}_0$	Transition probability kernel
$\mathcal{R}$	Set of all possible rewards
$\mathcal{X}$	State space
$\mathcal{X}^+$	Set of terminal states
π	Policy
r	Immediate reward function

1 Markov Decision Processes

1.1 Setting

A Markov decision process (MDP) models sequential decision making. The agent learns from the environment how to take actions to achieve a goal thorough interaction by maximizing the total reward it receives over time. Here we follow the notation in [1].

Formally, a MDP is a triplet $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$. Here $\mathcal{X}$ is the state space, $\mathcal{A}$ is the set of actions, and for each state x we denote by $\mathcal{A}(x) \subseteq \mathcal{A}$ the set of admissible actions from state x (for example, the set of all legals moves in a chess position). For simplicity we assume $\mathcal{A}(x) = \mathcal{A}$ for all x. Finally, $\mathcal{P}_0$ is the transition probability kernel that captures all information in the model:

$$\begin{aligned} \mathcal{P}_0 : \mathcal{X} \times \mathcal{A} &\to \mathrm{Prob}(\mathcal{X} \times \mathbb{R}) \\ (x, a) &\mapsto \mathcal{P}_0(\cdot \mid x, a) \end{aligned}$$

which maps (x, a) to a probability measure on the product space of states and rewards. In practice, the set of possible rewards $\mathcal{R} \subseteq \mathbb{R}$ is almost always a finite set.

The state transition probability kernel $\mathcal{P}$ is the first marginal of $\mathcal{P}_0$, which gives for $x, a, y \in \mathcal{X} \times \mathcal{A} \times \mathcal{X}$ the probability of moving from state x to state y via action a:

$$\mathcal{P}(x, a, y) := \mathcal{P}_0(\{y\} \times \mathbb{R} \mid x, a).$$

We also view $\mathcal{P}$ as a probability kernel that sends (x, a) to a distribution $\mathcal{P}(\cdot \mid x, a)$.

The immediate reward function r defined by

$$\begin{aligned} r : \mathcal{X} \times \mathcal{A} &\to \mathbb{R} \\ (x, a) &\mapsto \mathbb{E}[R_{(x,a)}] \end{aligned}$$

is the expectation of the second component of $\mathcal{P}_0$, where $(X_{(x,a)}, R_{(x,a)}) \sim \mathcal{P}_0(\cdot \mid x, a)$. Clearly we have

$$r(x, a) = \sum_{r \in \mathcal{R}} r \sum_{y \in \mathcal{X}} \mathcal{P}_0(y, r \mid x, a).$$

We impose the assumption that $\|r\|_\infty < \infty$.

The interaction of the agent with the environment is modelled as follows. At time $t \in \mathbb{N}$ (initially $t = 0$), the agent is at state $X_t \in X$ and chooses an action $A_t \in \mathcal{A}(X_t)$ based on some rules. The next state and the reward is given by

$$(X_{t+1}, R_{t+1}) \sim \mathcal{P}_0(\cdot \mid X_t, A_t).$$

The dynamic together with an initial random state X_0 gives rise to a sequence

$$X_0, A_0, R_1, \cdots, X_{t-1}, A_{t-1}, R_t, X_t,$$

called a trajectory of length t. A rule for selecting actions is called a behavior or a policy, which determines the distribution of trajectories. The goal is to find an optional policy, i.e. one that maximizes the expected return $\mathbb{E}[G]$ irrespective of the how the process started, where the return is defined by

$$G := \sum_{t=0}^{\infty} \gamma^t R_{t+1}$$

for some discount factor $\gamma \in [0, 1]$. We also define the future return from time t by

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

Clearly $G = G_0$ and we have the recursive relationship

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

The trajectory of an MDP in particular gives rise to a Markov chain on the state space $(X_t)_{t\in\mathbb{N}}$. Let $\mathcal{X}^+$ denote the set of all terminal (or absorbing) states of this chain. If $\mathcal{X}^+ \neq \varnothing$, the MDP is called episodic. In this case we often take $\gamma = 1$, and the future reward at every terminal state is set to 0. This means that the sum defining G is essentially a finite sum which stops at the random time T, the hitting time of $\mathcal{X}^+$. When $\mathcal{X}^+ = \varnothing$, we say that the MDP is continuing. In what follows, unless otherwise stated we assume that the MDP is continuing for notational convenience. Nevertheless, most discussion may still apply to the episodic case with minor adjustments.

Before going any further, let us rigorously define what a policy is. It is natural to model policies as a way of decision making that takes into the account the observed trajectory. Formally, a policy is a probability kernel π that maps any history $h_t := (x_0, a_0, r_1, ..., x_{t-1}, a_{t-1}, r_t, x_t)$ of length t to a probability distribution $\pi(\cdot \mid h_t)$ over $\mathcal{A}(x_t)$. If $H_t := (X_0, A_0, R_1, ..., X_{t-1}, A_{t-1}, R_t, X_t)$ then the next action $A_t \sim \pi(\cdot \mid H_t)$ is drawn according to this policy.

Stationary policies are a special type of policy where $\pi(\cdot \mid H_t)$ depends only on the current position X_t instead of the entire history H_t, making the behavior time-independent and hence the name stationary. In other words, a stationary policy is a mapping

$$\begin{aligned} \pi : \mathcal{X} &\to \operatorname{Prob}(\mathcal{A}) \\ x &\mapsto \pi(\cdot \mid x) \end{aligned}$$

that associates each state x with a distribution $\pi(\cdot \mid x)$ supported on $\mathcal{A}(x)$. This is a vast simplification while still being general enough (for example, chess evaluation can be made based solely on the current state of the game). In this case, the action

$$A_t \sim \pi(\cdot \mid X_t)$$

at time $t \in \mathbb{N}$ is now drawn from the distribution induced by the policy-state pair. A further simplification is to require the policy to be deterministic, namely a mapping $\pi : \mathcal{X} \to \mathcal{A}$. Clearly every stationary policy is a convex combination of deterministic stationary policies. Given an MDP, we denote the set of all policies by Π, the set of all stationary policies by Π_{stat}, and the set of all deterministic policies by Π_{det}. Due to the Markovian nature of $\mathcal{P}$, we restrict our attention to Π_{stat} unless otherwise stated.

Given an MDP $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$, a Markov reward process (MRP) $(\mathcal{X}, \mathcal{P}_0^\pi)$ is induced by fixing a stationary policy π. The name is changed from MDP to MRP because the decision-making aspect has been decided by the choice of π and hence the focus is solely on the rewards. Specifically, each state is assigned now the average probability distribution taken across actions:

$$\begin{aligned} \mathcal{P}_0^\pi : \mathcal{X} &\to \operatorname{Prob}(\mathcal{X} \times \mathbb{R}) \\ x &\mapsto \sum_{a \in \mathcal{A}} \pi(a \mid x) \mathcal{P}_0(\cdot \mid x, a). \end{aligned}$$

By the law of total expectation, if the current state is $X_t = x$, the expectation of the next reward R_{t+1} following policy π is

$$\begin{aligned} \mathbb{E}[R_{t+1} \mid X_t = x] &= \sum_{a \in \mathcal{A}(x)} \mathbb{P}(A_t = a \mid X_t = x) \mathbb{E}[R_{t+1} \mid X_t = x, A_t = a] \\ &= \sum_{a \in \mathcal{A}(x)} \pi(a \mid x) \sum_{r \in \mathcal{R}} r \sum_{y \in \mathcal{X}} \mathcal{P}_0(y, r \mid x, a). \end{aligned}$$

1.2 Value Functions

Let $\pi \in \Pi$ be a policy. To evaluate how good π is, we define its state-value function and action-value function.

The (state-)value function $V^\pi : \mathcal{X} \to \mathbb{R}$ for policy π is the expected return when the MDP starts from x and follows π thereafter; namely,

$$V^\pi(x) := \mathbb{E}[G \mid X_0 = x] = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} \middle| X_0 = x \right]$$

where X_0 is a distribution fully support on $\mathcal{X}$ so that the conditional expectation is well-defined. (Note that this definition is independent of the distribution of X_0). For an MRP induced by π we denote the value function by V for simplicity. Note that if $x \in \mathcal{X}^+$ then $V^\pi(x) = 0$.

The action-value function $Q^\pi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ for policy π is the expected return when the MDP starts from x, takes the action a and follows π thereafter; namely,

$$Q^\pi(x, a) := \mathbb{E}[G \mid X_0 = x \mid A_0 = a] = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} \middle| X_0 = x, A_0 = a \right],$$

where the joint distribution of X_0 and A_0 should be nonvanishing.

Recall that the (two-argument) immediate reward function can be written as

$$r(x,a) = \mathbb{E}[R_{t+1} \mid X_t = x, A_t = a].$$

For convenience we also consider the three-argument immediate reward function that gives the expected return given a state-action-next-state triplet:

$$\begin{aligned} r(x,a,y) &:= \mathbb{E}[R_{t+1} \mid X_t = x, A_t = a, X_{t+1} = y] \\ &= \sum_{r \in \mathcal{R}} r \frac{\mathbb{P}(R_{t+1} = r, X_{t+1} = y \mid X_t = x, A_t = a)}{\mathbb{P}(X_{t+1} = y \mid X_t = x, A_t = a)} \\ &= \sum_{r \in \mathcal{R}} r \frac{\mathcal{P}_0(y, r \mid x, a)}{\mathcal{P}(x,a,y)}. \end{aligned}$$

It is crucial that V^π and Q^π can be expressed in terms of each other. On the one hand,

$$V^\pi(x) = \mathbb{E}[G \mid X_0 = x] = \sum_{a \in \mathcal{A}} \mathbb{P}(A_0 = a \mid X_0 = x)\mathbb{E}[G \mid X_0 = x, A_0 = a],$$

so that

$$V^\pi(x) = \sum_{a \in \mathcal{A}} \pi(a \mid x) Q^\pi(x,a) = \mathbb{E}_{A \sim \pi(\cdot|x)}[Q^\pi(x,A)]. \tag{1}$$

On the other hand, since $G = R_1 + \gamma G_1$,

$$\begin{aligned} Q^\pi(x,a) &= \mathbb{E}[G \mid X_0 = x, A_0 = a] \\ &= \sum_y \mathbb{P}(X_1 = y \mid X_0 = x, A_0 = a)\mathbb{E}[G \mid X_0 = x, A_0 = a, X_1 = y] \\ &= \sum_y \mathcal{P}(x,a,y)\left(r(x,a,y) + \gamma \mathbb{E}[R_1 \mid X_0 = x, A_0 = a, X_1 = y]\right) \\ &= \sum_y \mathcal{P}(x,a,y)\left[r(x,a,y) + \gamma V^\pi(y)\right], \end{aligned}$$

where y is summed across the support of $\mathcal{P}(x,a,\cdot)$. Rewriting in terms of $\mathcal{P}_0$ yields

$$\begin{aligned} Q^\pi(x,a) &= \sum_y \sum_{r \in \mathcal{R}} \mathcal{P}_0(y, r \mid x, a)[r + \gamma V^\pi(y)] = r(x,a) + \gamma \sum_y \mathcal{P}(x,a,y) V^\pi(y) \\ &= r(x,a) + \gamma \mathbb{E}_{Y \sim \mathcal{P}(\cdot|x,a)} V^\pi(Y). \end{aligned} \tag{2}$$

The interpretation of these two equations is intuitive. Equation (1) says that the state value at x is the average of the state-action values, taken across all possible actions. Equation (2) says that the state-action value at (x,a) is the average of the next-step reward r and the discounted state value at y, taken across all possible rewards r and successor states y. Substituting the two equations into each other, we get what is called Bellman equations for V^π and Q^π, which are self-referential equations that expresses the value at a point to the average value across all successor states:

$$V^\pi(x) = \sum_{a,y,r} \pi(a \mid x)\mathcal{P}_0(y, r \mid x, a)[r + \gamma V^\pi(y)]; \tag{3}$$

$$Q^\pi(x, a) = \sum_{a',y,r} \pi(a' \mid y)\mathcal{P}_0(y, r \mid x, a)[r + \gamma Q^\pi(y, a')]. \tag{4}$$

Again, these two equations expresses the value of V^π and Q^π at a state or a state-action pair as the average of the values at their successors.

Note that Equations (3) and (4) are systems of linear equations and can be expressed in matrix form, using the following notation:

- View $V^\pi \in \mathbb{R}^{|\mathcal{X}|}$, $Q^\pi \in \mathbb{R}^{|\mathcal{X}||\mathcal{A}|}$, and $r \in \mathbb{R}^{|\mathcal{X}||\mathcal{A}|}$ as vectors.
- Let $r^\pi(x) = \sum_a \pi(a \mid x)r(x, a)$ and view it as a vector $r^\pi \in \mathbb{R}^{|\mathcal{X}|}$
- View $\mathcal{P}$ as a matrix of size $(|\mathcal{X}||\mathcal{A}|) \times |\mathcal{X}|$ where $\mathcal{P}_{(x,a),y} := \mathcal{P}(x, a, y)$.
- π induces transition matrices P^π on $\mathcal{X}$ and $\mathcal{P}^\pi$ on $\mathcal{X} \times \mathcal{A}$:

$$P^\pi_{x,y} := \sum_a \pi(a \mid x)\mathcal{P}(x, a, y)$$

$$\mathcal{P}^\pi_{(x,a),(y,a')} := \mathcal{P}(x, a, y)\pi(a' \mid y).$$

With these notations, equations (2)-(4) become

$$Q^\pi = r + \gamma\mathcal{P}V^\pi,\ \ V^\pi = r^\pi + \gamma P^\pi V^\pi,\ \ Q^\pi = r + \gamma\mathcal{P}^\pi Q^\pi.$$

When $0 \leq \gamma < 1$ for any nonzero $x \in \mathbb{R}^{|\mathcal{X}||\mathcal{A}|}$ we have

$$\|(I - \gamma\mathcal{P}^\pi)x\|_\infty \geq \|x\|_\infty - \gamma \|\mathcal{P}^\pi x\|_\infty \geq \|x\|_\infty - \gamma \|x\|_\infty > 0,$$

so $I - \gamma\mathcal{P}^\pi$ is invertible. Similarly, $I - \gamma P^\pi$ is invertible, and therefore

$$V^\pi = (I - \gamma P^\pi)^{-1}r^\pi, Q^\pi = (I - \gamma\mathcal{P}^\pi)^{-1}r.$$

Thus the policy evaluation for discounted finite MDPs can be solved as a linear system. In practice, however, this is not feasible beyond small instances. A more practical approach to solve for the value functions makes use of the Banach Fixed Point Theorem.

Theorem 1.1 (Banach Fixed Point Theorem)**.** *Let (X, d) be a complete metric space. If $T : X \to X$ is a contraction, namely for all $x, y \in X$ we have $d(Tx, Ty) \leq \gamma d(x, y)$ where $0 \leq \gamma < 1$, then there exists a unique x^* such that $Tx^* = x^*$. Moreover, the rate of convergence is geometric: For any $x_0 \in X$ define $x_n := T^n x_0$. Then*

$$d(x_n, x^*) \leq \frac{\gamma^n}{1 - \gamma}d(x_1, x_0).$$

In particular, the fixed point can be found by iteratively applying T starting from any initial point.

For a set X let $\mathcal{B}(X)$ be the normed space of all $\mathbb{R}$-valued bounded functions on X, equipped with the sup-norm $\|\cdot\|_\infty$. It is easy to see that $\mathcal{B}(X)$ is a Banach space. This is because a Cauchy sequence in $\mathcal{B}(X)$ is a uniformly Cauchy sequence of functions from X to $\mathbb{R}$, and since $\mathbb{R}$ is complete, this sequence of function would converge uniformly to a limit, which is precisely convergence with respect to $\|\cdot\|_\infty$. In particular, $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\mathcal{X}\times\mathcal{A})$ are complete.

Next, for a fixed policy π we define the Bellman operators $T_V^\pi : \mathcal{B}(\mathcal{X}) \to \mathcal{B}(\mathcal{X})$ and $T_Q^\pi : \mathcal{B}(\mathcal{X}\times\mathcal{A}) \to \mathcal{B}(\mathcal{X}\times\mathcal{A})$ by

$$\begin{aligned}(T_V^\pi V_0)(x) &:= \sum_{a,y,r} \pi(a \mid x)\mathcal{P}_0(y, r \mid x, a)[r + \gamma V_0(y)];\\ (T_Q^\pi Q_0)(x,a) &:= \sum_{a',y,r} \pi(a' \mid y)\mathcal{P}_0(y, r \mid x, a)[r + \gamma Q_0(y,a')],\end{aligned}$$

or in matrix form,

$$\begin{aligned}T_V^\pi &: V_0 \mapsto r^\pi + \gamma P^\pi V_0;\\ T_Q^\pi &: Q_0 \mapsto r + \gamma \mathcal{P}^\pi Q_0.\end{aligned}$$

These operators are well-defined provided that the immediate reward function is bounded. By the Bellman equations (3) and (4), V^π and Q^π are fixed points of T_V^π and T_Q^π respectively, if they exist. Observe that T_V^π and T_Q^π are γ-Lipschitz:

$$\|T_V^\pi V_1 - T_V^\pi V_2\|_\infty \le \gamma \sup_{x\in\mathcal{X}} \sum_{a,y,r} \pi(a \mid x)\mathcal{P}_0(y, r \mid x, a)|V_1(y) - V_2(y)| \le \gamma \|V_1 - V_2\|_\infty$$

and similarly for T_Q^π. When $0 \le \gamma < 1$, the existence and uniqueness of V^π and Q^π are guaranteed by Banach Fixed Point Theorem. Therefore, we can initialize $V := V_0$ at will and iteratively apply the operator $V_{k+1} := T_V^\pi V_k$, and V_k will converge to V^π in geometric speed (and similarly for Q^π). This procedure is called iterative policy evaluation.

1.3 Optimal Value Functions

The optimal (state-)value function $V^\star : \mathcal{X} \to \mathcal{X}$ and the optimal action-value function $Q^\star : \mathcal{X}\times\mathcal{A} \to \mathcal{X}\times\mathcal{A}$ are defined as the best possible state and state-action values:

$$V^\star(x) := \sup_{\pi\in\Pi} V^\pi(x), \;\; Q^\star(x,a) := \sup_{\pi\in\Pi} Q^\pi(x,a).$$

We consider a natural pre-order on Π: $\pi \le \pi'$ if $V^\pi(x) \le V^{\pi'}(x)$ for all $x \in \mathcal{X}$. We say that a policy $\pi^\star \in \Pi$ is optimal if for all $\pi \in \Pi$ we have $\pi \le \pi^\star$ (note that this is not a partial order because antisymmetry does not hold). It is not at all clear why $\pi^\star$ exists, because the maximizing policy in $V^\star$ for each x may be different. At the very least, by definition we have $V^{\pi_1} = V^{\pi_2}$ for any optimal policies $\pi_1, \pi_2 \in \Pi$. It turns out that for finite MDPs there exist optimal policies $\pi \in \Pi_{\text{det}}$ that are stationary and deterministic. In what follows, we first prove the existence of optimal policies [2], and then discuss how to find one using greediness.

Theorem 1.2 (Existence of Optimal Policies)**.** *Given a countable MDP with a finite set of actions, there exists some stationary deterministic policy $\pi \in \Pi_{det}$ which is optimal. Moreover, any optimal policy $\pi \in \Pi$ satisfies $Q^{\pi}(x, a) = Q^{\star}(x, a)$ for all $(x, a) \in \mathcal{X} \times \mathcal{A}$.*

Proof. The plan of the proof is as follows. We first show that given any policy, the maximum expected reward does not depend on the entire history h_t but on the current state x_t only. This allows us to define a greedy action in terms of $V^{\star}$ which we will show to be optimal.

Specifically, we start by showing that for any $(x, a, r, y) \in \mathcal{X} \times \mathcal{A} \times \mathcal{R} \times \mathcal{X}$ we have

$$\sup_{\pi \in \Pi} \mathbb{E}_{\pi} \left[\sum_{t=1}^{\infty} \gamma^t R_{t+1} \middle| (X_0, A_0, R_0, X_1) = (x, a, r, y) \right] = \gamma V^{\star}(y).$$

For $\pi \in \Pi$ and $(x, a, r) \in \mathcal{X} \times \mathcal{A} \times \mathcal{R}$ define the policy

$$\pi_{(x,a,r)} : h_t \mapsto \pi(\cdot \mid x, a, r, h_t)$$

that makes decisions by prepending (x, a, r) to any history h_t. Note that for any triple (x, a, r) we have $\Pi = \{\pi_{(x,a,r)} : \pi \in \Pi\}$: for $\pi \in \Pi$ consider the policy

$$\pi_{-1} : h_t = (x_0, a_0, r_0, \tau) \mapsto \pi_{-1}(\cdot \mid h_t) := \pi(\cdot \mid \tau)$$

that "forgets" the first segment of every trajectory, so that $\pi = (\pi_{-1})_{(x,a,r)}$. With this notation, we can use the Markov property and see that

$$\mathbb{E}_{\pi} \left[\sum_{t=1}^{\infty} \gamma^t R_{t+1} \middle| (X_0, A_0, R_0, X_1) = (x, a, r, y) \right] = \gamma \mathbb{E}_{\pi_{(x,a,r)}}[G_1 \mid X_1 = y] = \gamma V^{\pi_{(x,a,r)}}(y),$$

and taking the supremum establishes the equality.

Next, we show that the policy $\tilde{\pi} \in \Pi_{\text{det}}$ defined by

$$\tilde{\pi}(x) := \arg\max_{a \in \mathcal{A}} \left[r(x, a) + \gamma \mathbb{E}_{y \sim \mathcal{P}(x,a)} V^{\star}(y) \right]$$

is optimal (in case when there are multiple maximizing actions, ties are broken arbitrarily). Indeed, for $x_0 \in \mathcal{X}$ we have

$$\begin{aligned}
V^\star(x_0) &= \sup_{\pi\in\Pi} \mathbb{E}_\pi \left[r(x_0, A_0) + \sum_{t=1}^{\infty} \gamma^t r(X_t, A_t) \right] \\
&= \sup_{\pi\in\Pi} \mathbb{E}_\pi \left[r(x_0, A_0) + \mathbb{E}_\pi \left[\sum_{t=1}^{\infty} \gamma^t r(X_t, A_t) \middle| X_0 = x_0, A_0, R_0, X_1 \right] \right] \\
&\le \sup_{\pi\in\Pi} \mathbb{E}_\pi \left[r(x_0, A_0) + \sup_{\pi'\in\Pi} \mathbb{E}_{\pi'} \left[\sum_{t=1}^{\infty} \gamma^t r(X_t, A_t) \middle| X_0 = x_0, A_0, R_0, X_1 \right] \right] \\
&= \sup_{\pi\in\Pi} \mathbb{E}_\pi \left[r(x_0, A_0) + \gamma V^\star(X_1) \right] \\
&\le \sup_{a_0\in\mathcal{A}} \left[r(x_0, a_0) + \gamma \mathbb{E}_{y\sim\mathcal{P}(x_0,a_0)} V^\star(y) \right] = \mathbb{E}_{\tilde\pi}[R_1 + \gamma V^\star(X_1)],
\end{aligned}$$

where X_1 and R_1 (and hence the entire trajectory) are almost surely constant because $\tilde\pi \in \Pi_{\text{det}}$. Note that the last inequality is in fact an equality because the supremum across Π is attained by $\tilde\pi$. Repeatedly applying this inequality yields

$$V^\star(x_0) \le \mathbb{E}_{\tilde\pi}[R_1 + \gamma V^\star(X_1)] \le \mathbb{E}_{\tilde\pi}[R_1 + \gamma R_2 + \gamma^2 V^\star(X_2)] \le \cdots \le V^{\tilde\pi}(x_0),$$

where we used dominated convergence theorem and the assumption that the reward function (and hence $V^\star$) is bounded. Thus $V^\star \le V^{\tilde\pi}$. Since $V^{\tilde\pi} \le V^\star$ trivially, we have $V^{\tilde\pi} = V^\star$ and $\tilde\pi$ is optimal.

Finally, we show that $Q^\star = Q^{\pi^\star}$ for any optimal policy $\pi^\star \in \Pi$. Trivially we have $Q^\star \ge Q^{\pi^\star}$. As for the other direction, by equation (2) and the definition of $Q^\star$ we have

$$Q^\star(x,a) = r(x,a) + \sup_{\pi\in\Pi} \mathbb{E}_{y\sim\mathcal{P}(\cdot|x,a)} V^\pi(y) \le r(x,a) + \mathbb{E}_{y\sim\mathcal{P}(\cdot|x,a)} V^\star(y) = Q^{\pi^\star}(x,a),$$

where we used $V^{\pi^\star} = V^\star$ in the last equality. □

Now that we establish the existence of an optimal policy, it remains to solve for the optimal action-value function and find an optimal policy. To start with, since there exists maximizing policy for value function, we should have a version of equations (1) and (2) for optimal value functions. Indeed, for all $x \in \mathcal{X}$ we have

$$V^\star(x) = \max_{a\in\mathcal{A}} Q^\star(x,a) \tag{5}$$

To see this, let $\pi^\star \in \Pi_{\text{det}}$ be an optimal policy. For any $a \in \mathcal{A}$ let π' be the policy that first takes action a and then follows $\pi^\star$. Then $Q^\star(x,a) = V^{\pi'}(x) \le V^\star(x)$, and since a is arbitrary we have $V^\star(x) \ge \max_{a\in\mathcal{A}} Q^\star(x,a)$. As for the other direction,

$$V^\star(x) = \sup_{\pi\in\Pi} \mathbb{E}_{a\sim\pi(\cdot|x)} Q^\pi(x,a) \le \sup_{\pi\in\Pi} \mathbb{E}_{a\sim\pi(\cdot|x)} Q^\star(x,a) = \max_{a\in\mathcal{A}} Q^\star(x,a).$$

Next, using equation 4 and the fact that $V^\star = V^{\pi^\star}$ and $Q^\star = Q^{\pi^\star}$, we have

$$Q^\star(x,a) = r(x,a) + \gamma \mathbb{E}_{x' \sim \mathcal{P}(\cdot|x,a)} V^\star(x'). \tag{6}$$

Therefore, we have the Bellman optimality equations

$$V^\star(x) = \max_{a \in \mathcal{A}} \left[r(x,a) + \gamma \mathbb{E}_{x' \sim \mathcal{P}(\cdot|x,a)} V^\star(x') \right] \tag{7}$$

as well as

$$Q^\star(x,a) = r(x,a) + \gamma \mathbb{E}_{x' \sim \mathcal{P}(\cdot|x,a)} \max_{a' \in \mathcal{A}} Q^\star(x',a') \tag{8}$$

for any $x \in \mathcal{X}$ and $a \in \mathcal{A}$. The one-step look-ahead explanation naturally carries over to equations (5)-(8).

Analogous to what we did in the policy evaluation section, we define the operators $T_V^\star : \mathcal{B}(\mathcal{X}) \to \mathcal{B}(\mathcal{X})$ and $T_Q^\star : \mathcal{B}(\mathcal{X} \times \mathcal{A}) \to \mathcal{B}(\mathcal{X} \times \mathcal{A})$ by

$$(T_V^\star V)(x) := \max_{a \in \mathcal{A}} \left[r(x,a) + \gamma \mathbb{E}_{x' \sim \mathcal{P}(\cdot|x,a)} V(x') \right] ;$$
$$(T_Q^\star Q)(x,a) := r(x,a) + \gamma \mathbb{E}_{x' \sim \mathcal{P}(\cdot|x,a)} \max_{a' \in \mathcal{A}} Q(x',a'),$$

but this time both operators are nonlinear. We say that V (resp. Q) satisfies the Bellman optimality equation if it is a fixed point of $T_V^\star$ (resp. $T_Q^\star$). It is straightforward to see that both $T_V^\star$ and $T_Q^\star$ are γ-Lipschitz, using the fact that

$$|\max_t f(t) - \max_t g(t)| \leq \max_t |f(t) - g(t)|.$$

Thus when $0 \leq \gamma < 1$, $T_Q^\star$ (resp. $T_V^\star$) admits a unique fixed point $Q^\star$ (resp. $V^\star$).

Now that we are able to solve for $Q^\star := \lim_{k \to \infty} (T_Q^\star)^k Q_0$ by iteration, the final task is to find a policy π such that $Q^\pi = Q^\star$. As the proof of Theorem 1.2 suggests, we should look at greedy policies. Specifically, for a action-value function Q resulted from some given policy, define the greedy policy $\pi_Q \in \Pi_{\mathrm{det}}$ with respect to Q by

$$\pi_Q(x) := \arg\max_{a \in \mathcal{A}} Q(x,a),$$

where the tie is broken arbitrarily if $\max_{a \in \mathcal{A}} Q(x,a)$ is attained by multiple elements.

Theorem 1.3. *For any action value function Q, we have that $Q = Q^\star$ if and only if Q satisfies the Bellman optimality equation. Furthermore, the greedy action $\pi_{Q^\star}$ with respect to $Q^\star$ is an optimal policy.*

Proof. The first assertion follows from equation (8) and the uniqueness of fixed points of $T_Q^\star$. To show that $\pi := \pi_{Q^\star}$ is optimal, we show that $Q^\pi = Q^\star$, and by the first assertion it suffices to show that Q^π satisfies the Bellman optimality equation. This directly follows from the Bellman equation for Q^π and the definition of $\pi_{Q^\star}$:

$$Q^\pi(x,a) = r(x,a) + \mathbb{E}_{\substack{x' \sim \mathcal{P}(\cdot|x,a) \\ a' \sim \pi(\cdot|x')}} Q^\pi(x',a') = r(x,a) + \mathbb{E}_{x' \sim \mathcal{P}(\cdot|x,a)} \max_{a' \in \mathcal{A}} Q^\pi(x',a')$$

for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$. □

1.4 Dynamic Programming for Solving MDPs

In reinforcement learning one is interested in solving two types of problems: prediction problem (evaluating the value function for a given policy) and more importantly the control problem (finding the optimal policy and/or optimal state-action value function). In this section we describe two ways to solve for MDPs, the value iteration algorithm and the policy iteration algorithms.

1.4.1 Value Iteration

We have seen value iteration in the previous section. One starts with an initialized action value function Q_0 and repeatedly apply the Bellman optimality operator

$$Q_{k+1} := T^\star_Q Q_k,$$

so that $Q_k \to Q^\star$ at a geometric rate. (Alternatively, one can work with V_k.) Once an action-value function Q is close to optimal, the associated greedy policy π_Q is also close to optimal, with error

$$V^{\pi_Q}(x) \geq V^\star(x) - \frac{2}{1-\gamma} \left\|Q - Q^\star\right\|_\infty$$

for all state x. To see this, let $\pi^\star \in \Pi_{\text{det}}$ be optimal. Then

$$\begin{aligned} V^\star(x) - V^{\pi_Q}(x) &= Q^\star(x, \pi^\star(x)) - Q^{\pi_Q}(x, \pi_Q(x)) \\ &= [Q^\star(x, \pi^\star(x)) - Q(x, \pi_Q(x))] + [Q(x, \pi_Q(x)) - Q^\star(x, \pi_Q(x))] + \\ &\quad [Q^\star(x, \pi_Q(x)) - Q^{\pi_Q}(x, \pi_Q(x))]. \end{aligned}$$

Since π_Q is greedy, the first bracket is bounded by

$$Q^\star(x, \pi^\star(x)) - Q(x, \pi_Q(x)) \leq Q^\star(x, \pi^\star(x)) - Q(x, \pi^\star(x)) \leq \left\|Q^\star - Q\right\|_\infty.$$

The second bracket is also bounded by $\left\|Q^\star - Q\right\|_\infty$. By the Bellman equations, the third bracket is bounded by

$$Q^\star(x, \pi_Q(x)) - Q^{\pi_Q}(x, \pi_Q(x)) = \gamma \mathbb{E}_{y \sim \mathcal{P}(\cdot|x, \pi_Q(x))}[V^\star(y) - V^{\pi_Q}(y)] \leq \gamma \left\|V^\star - V^{\pi_Q}\right\|_\infty.$$

Combining these three estimates yields the result.

1.4.2 Policy Iteration

The policy iteration algorithm works by alternating between policy evaluation and policy improvement. Specifically, initialize a policy π_0. At each step $k \geq 0$, evaluate the value function Q^{π_k} (or V^{π_k}) using the iterative method given by the Bellman equation (see the end of Section 1.2). After we have computed Q^{π_k}, we then perform the so-called policy improvement step by choosing the next policy $\pi_{k+1} := \pi_{Q^{\pi_k}}$ as greedy with respect to Q^{π_k}. This is indeed an improvement due to the following theorem [3].

Theorem 1.4 (Policy Improvement Theorem)**.** *Let $\pi, \pi' \in \Pi_{det}$ be deterministic policies. If $Q^\pi(x, \pi'(x)) \geq V^\pi(x)$ for all $x \in \mathcal{X}$, then $\pi' \geq \pi$. Moreoever,*

1. *for any $x \in \mathcal{X}$ such that $Q^\pi(x, \pi'(x)) > V^\pi(x)$ we have $V^{\pi'}(x) > V^\pi(x)$;*
2. *if $Q^\pi(x, \pi'(x)) = V^\pi(x)$ for all $x \in \mathcal{X}$ we have $V^{\pi'} = V^\pi$, and if furthermore $\pi' = \pi_{Q^\pi}$ is greedy with respect to Q^π, then both π and π' are optimal.*

Proof. By the assumption and equation (2),

$$\begin{aligned} V^\pi(x) \leq Q^\pi(x, \pi'(x)) &= \mathbb{E}_\pi[R_1 + \gamma V^\pi(X_1) \mid X_0 = x, A_0 = \pi'(x)] \\ &= \mathbb{E}_{\pi'}[R_1 + \gamma V^\pi(X_1) \mid X_0 = x] \\ &\leq \mathbb{E}_{\pi'}[R_1 + \gamma Q^\pi(X_1, \pi'(X_1)) \mid X_0 = x] \\ &= \mathbb{E}_{\pi'}[R_1 + \gamma \mathbb{E}_\pi[R_2 + \gamma V^\pi(X_2) \mid X_1, A_1 = \pi'(X_1)] \mid X_0 = x] \\ &= \mathbb{E}_{\pi'}[R_1 + \gamma \mathbb{E}_{\pi'}[R_2 + \gamma V^\pi(X_2) \mid X_1] \mid X_0 = x] \\ &= \mathbb{E}_{\pi'}[R_1 + \gamma R_2 + \gamma^2 V^\pi(X_2) \mid X_0 = x], \end{aligned}$$

and repeatedly applying this argument gives for all $k \in \mathbb{N}$

$$V^\pi(x) \leq \mathbb{E}_{\pi'}\left[\sum_{j=0}^{k-1} \gamma^j R_{j+1} + \gamma^k V^\pi(X_k)\right] \to V^{\pi'}(x) \text{ as } k \to \infty,$$

where we once again used dominated convergence theorem and the boundedness assumption on the reward.

The assertions about the strict inequality and equality between V^π and $V^{\pi'}$ directly follow from the same argument. As for the case when the greedy policy π' with respect to Q^π does not improve upon π, for every $x \in \mathcal{X}$ we have

$$V^\pi(x) = Q^\pi(x, \pi'(x)) = \max_{a \in \mathcal{A}} Q^\pi(x, a) = \max_{a \in \mathcal{A}} [r(x, a) + \gamma \mathbb{E}_{y \sim \mathcal{P}(\cdot|x,a)} V^\pi(y)].$$

In other words, V^π satisfies the Bellman optimality equation (7), so π (and hence π') are optimal. □

Remark 1.1. If we allow $\pi, \pi' \in \Pi_{\text{stat}}$ and define

$$Q(x, \pi'(x)) := \mathbb{E}_{a \sim \pi'(\cdot|x)} Q(x, a),$$

then it is still true that $Q^\pi(x, \pi'(x)) \geq V^\pi(x)$ for all $x \in \mathcal{X}$ implies $\pi' \geq \pi$. To see this, simply replace parts $A_t = \pi'(X_t)$ in the proof with $A_t \sim \pi'(\cdot \mid X_t)$. The analogy of 2. can be made precise by the notion of ϵ-greedy policies (see Proposition 2.1).

By taking π' to be greedy with respect to Q^π, we always have $Q^\pi(x, \pi'(x)) \geq V^\pi(x)$ by equation (1). Therefore during the policy iteration algorithm, at each stage the policy will strictly improve at some state until it reaches an optimal policy. Regarding the convergence of policy iteration, we have the following result.

Theorem 1.5 (Policy Iteration)**.** *Fixed an initial policy π_0 and for $k \geq 0$ let $Q_k := Q^{\pi_k}$ and $\pi_{k+1} := \pi_{Q^{\pi_k}}$. Then $Q_{k+1} \geq T_Q^\star Q_k \geq Q_k$ and $\|Q^\star - Q_{k+1}\|_\infty \leq \gamma \|Q^\star - Q_k\|_\infty$.*

Proof. From the policy improvement theorem we know that $Q_{k+1} \geq Q_k$. Thus

$$\begin{aligned} Q_{k+1}(x,a) &= r(x,a) + \gamma \mathbb{E}_{y \sim \mathcal{P}(\cdot|x,a)} Q_{k+1}(y, \pi_{k+1}(y)) \\ &\geq r(x,a) + \gamma \mathbb{E}_{y \sim \mathcal{P}(\cdot|x,a)} Q_k(y, \pi_{k+1}(y)) \\ &= r(x,a) + \gamma \mathbb{E}_{y \sim \mathcal{P}(\cdot|x,a)} \max_{a' \in \mathcal{A}} Q_k(y, a') = T_Q^\star Q_k(x,a) \\ &\geq r(x,a) + \gamma \mathbb{E}_{y \sim \mathcal{P}(\cdot|x,a)} Q_k(y, \pi_k(y)) = Q_k(x,a), \end{aligned}$$

and since $Q_{k+1} \geq T_Q^\star Q_k$ we have

$$\|Q^\star - Q_{k+1}\|_\infty \geq \left\|Q^\star - T_Q^\star Q_k\right\|_\infty = \left\|T_Q^\star Q^\star - T_Q^\star Q_k\right\|_\infty \leq \gamma \|Q^\star - Q_k\|_\infty .$$

□

How does value iteration and policy iteration compare? Let Q_0 be any initial action-value function. Denote the k-th action value function in value iteration and policy evaluation by Q_k^v and Q_k^p respectively, namely

$$Q_k^v := T_Q^\star Q_{k-1}^v, \;\; Q_k^p := Q^{\pi_{Q_{k-1}^p}}.$$

The policy iteration theorem tells us

$$Q_1^p \geq Q_1^v, \; Q_2^p \geq T_Q^\star Q_1^p \geq T_Q^\star Q_1^v = Q_2^v, \; Q_3^p \geq T_Q^\star Q_2^p \geq T_Q^\star Q_2^v = Q_3^v, ...$$

and thus policy iteration gives better action value function than value iteration after the same number of iterations. However, the policy iteration is much more computationally intensive as each iteration involves a policy evaluation step.

Almost all control problems in reinforcement learning are solved via some sort of generalized policy iteration, where one alternate between policy evaluation and policy improvement without requiring perfect convergence at each step. As long as all states are being updated, convergence result can typically be obtained at various levels of granularity.

1.5 Finite-Horizon MDPs

Finite-horizon Markov decision processes form a subclass of episodic MDPs and is typically more amenable to analysis. Formally, a finite-horizon MDP is given by a tuple $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_h, r_h, H)$, where $\mathcal{X}$ is the state space, $\mathcal{A}$ is the action space, $H \in \mathbb{N}$ is the horizon which specifies the episode length of this MDP, and $\mathcal{P}_h : \mathcal{X} \times \mathcal{A} \to \text{Prob}(\mathcal{X})$ and $r_h : \mathcal{X} \times \mathcal{A} \to [0,1]$ are the time-dependent transition dynamics and immediate reward functions ($h = 1, ..., H-1$). In other words, the length of each episode in this MDP is H, and at each step $h = 1, ..., H-1$ the dynamic evolves according to $\mathcal{P}_h$

and generates immediate reward r_h (assumed to be bounded in $[0,1]$ without loss of generality). We define the return associated with a trajectory to ve the undiscounted sum of immediate rewards collected along the way.

Let π be a (possibly non-stationary and randomized) policy. For $(x,a,h) \in \mathcal{X} \times \mathcal{A} \times [H]$ we similarly define the state-value function (at time h following π) as

$$V_h^\pi(x) := \mathbb{E}_\pi \left[\sum_{t=h}^{H-1} r_h(x_t, a_t) \middle| x_h = x \right]$$

as well as the action-value function

$$Q_h^\pi(x,a) := \mathbb{E}_\pi \left[\sum_{t=h}^{H-1} r_h(x_t, a_t) \middle| x_h = x, a_h = a \right],$$

where x_t and a_t form a trajectory of the MRP induced by π. The optimal value function is defined as the supremum of $V^\pi := V_0^\pi$ across all policies π, and a policy is called optimal if it attains this supremum.

We can view a finite-horizon MDP as a stationary MDP (i.e. our original setting) by incorporating the time parameter $h \in [H]$ as part of the state. Namely we can consider the induced MDP with state space $\mathcal{X} \times [H]$, and at each step $h \in [H-1]$ the current state is in $\mathcal{X} \times [h]$, transitions to $\mathcal{X} \times [h+1]$ following $\mathcal{P}_h$ and collects reward according to r_h. Then it is an immediate corollary of Theorem 1.2 that optimal policies can be chosen to be of the form $\pi : (x,h) \in \mathcal{X} \times [H-1] \mapsto \mathcal{A}$ which depends only on the state and the time parameter. The time dependence is generally necessary, as illustrated by the simple MDP with $\mathcal{X} = \{x_0\}$, $H = 3$, $\mathcal{A} = \{a, b\}$ (both leading back to x_0) and $r_1(s, a_1) = r_2(s, a_2) = 1$, $r_1(s, a_2) = r_2(s, a_1) = 0$, for which the optimal policy is $\pi(s,h) = a_h$ $(h = 1, 2)$.

Moreover, the same "feed-forward" perspective allows us to obtain the Bellman optimality equations for finite-horizon MDPs. Define

$$Q_h^\star(x,a) := \sup_{\pi \in \Pi} Q_h^\pi(x,a)$$

and for all π let $Q_H^\pi = 0$ by convention. Then counterparts of Equation (4) and (8) are

$$Q_h^\pi(x,a) = r_h(x,a) + \mathbb{E}_{x' \sim \mathcal{P}_h(\cdot|x,a), a' \sim \pi(s,h)} Q_{h+1}^\pi(x', a'); \tag{9}$$

$$Q_h^\star(x,a) = r_h(x,a) + \mathbb{E}_{x' \sim \mathcal{P}_h(\cdot|x,a)} \max_{a' \in \mathcal{A}} Q_{h+1}^\star(x', a'). \tag{10}$$

Let $Q_h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ $(h \in [H])$ be a sequence of mappings with $Q_H = 0$. Then $\{Q_h\}_h$ is the optimal action-value function if and only if $\{Q_h\}_h$ (when substituted into $\{Q_h^\star\}_h$) satisfies (10) . The "only if" part follows from (8), while the "if" part is true because any $\{Q_h\}_h$ with $Q_H = 0$ satisfying (10) is uniquely determined: $Q_{H-1} = r_{H-1}$ and Q_{H-K} is completely determined by r_{H-k} $(k = 1,..,K)$ and $\mathcal{P}_{H-k}$ $(k = 1,..,K-1)$. Moreover, the greedy action $\pi(x,h) := \arg\max_{a \in \mathcal{A}} Q_h^\star(x,a)$ is optimal, since its value function satisfies (10).

2 Monte Carlo Methods

In the first chapter we discussed MDPs and dynamic programming algorithms to solve for them. These methods fall into the category of model-based approaches, where the word model refers to the probability kernel $\mathcal{P}$. In practice, however, it is often intractable or infeasible to model a problem as an MDP and compute all the transition probabilities. In such cases, model-free approaches such as temporal difference (TD) and Monte Carlo would be in favor, which relies only on experience (i.e. sampled trajectories) instead of a model. In TD and Monte Carlo, we seek to estimate the state-action value function, since state-value function alone is insufficient to determine an optimal policy (e.g. greedy) in the absence of a model. In this section we discuss Monte Carlo Methods

The spirit of MC methods is to estimate $V^\pi(x) = \mathbb{E}_\pi[G_t \mid X_t = x]$ and $Q^\pi(x, a) = \mathbb{E}_\pi[G_t \mid X_t = x, A_t = a]$ by averaging the returns observed after visiting a state x or a state-action pair (x, a) in simulated trajectories. We focus on episodic MDP to ensure that the return is well-defined, since for continuing tasks the return is defined by an infinite sum and the exact data is impossible to collect.

2.1 Monte Carlo Prediction

Fix a policy π to be evaluated. The definition of V^π suggests that we generate episodes

$$X_0, A_0, R_1, X_1, A_1, R_2, ..., X_{T-1}, A_{T-1}, R_T$$

following π and then estimate $V^\pi(x)$ (resp. $Q^\pi(x, a)$) as the simple average of G_t for t such that $X_t = x$ (resp. $(X_t, A_t) = (x, a)$). If we average G_t's for all such t we get what is called the *every-visit MC method*, while if we only take G_t's such that each episode first encounters the state x (resp. the state-action pair (x, a)) at time t, we get what is called the *first-visit MC method*. Every-visit MC has the computational advantage that one does not need to keep track of which states have been visited, but analysis for first-visit MC is simpler because estimates for return at one state are iid.

In first-visit MC, for every state x (resp. state-action pair (x, a)) the empirical returns G_t (across episodes) are iid with expectation $V^\pi(x)$ (resp. $Q^\pi(x, a)$); in other words, first-visit MC method gives unbiased estimates. Thus by the law of large numbers, the estimates indeed converge to the true value functions, and the variance of the estimates is of order $1/n$ where n is the number of returns averaged.

Every-visit MC is more complicated [4]. Let's say we want to estimate $V^\pi(x)$ (the case for $Q^\pi(x, a)$ is similar). We can simplify the Markov chain on $\mathcal{X}$ induced by the MRP to a two-state chain, one being x and the other being the collection of all terminal states. Each episode starting from x can be decomposed into segments which start and end at x and the last segment that starts at x and ends at some terminal state. This point of view defines a MRP on the two-state chain and can be used to analyze every-visit MC. In fact, every-visit MC is biased (where the bias is of order

$1/n$, so it is asymptotically unbiased) and the estimates have variance of order $1/n$ where n is the number of episodes.

One major issue with MC methods is that each state (or state-action pair) needs to be visited infinitely often (sufficiently many times in practice) for convergence to happen, which is not possible when the state(-action) space is large. One simply workaround is impose the assumption of *exploring starts*, which requires that every state-action pair has a non-zero probability of being selected at the start. However, this assumption is unlikely to hold when directly learning from actual interaction with the environment. The most common alternative is to consider policies that has a nonzero probability of selection all actions in each state (what we call *soft* policies). Another alternative is *off-policy learning*, in which one separates the policy used to generate trajectories (called the *behavior policy*) from the policy being optimized (called the *target policy*). This way, one can achieve better exploration by selecting a more stochastic behavior policy while exploiting the past experience with the target policy (which is often greedy). In what follows we discuss these three methods: Monte Carlo with exploring starts, on-policy MC for ϵ-soft policies, and off-policy MC.

2.2 On-Policy Monte Carlo Control

On-policy learning refers to RL algorithms in which the behavior policy coincide with the target policy, as opposed to off-policy learning.

2.2.1 Monte Carlo with Exploring Starts (MCES)

Monte Carlo with Exploring Starts (MCES) is a version of GPI that combines MC prediction described above and policy improvement via greedy policies, together with the exploring starts assumption. This approach is the same in spirit as GPI in dynamic programming, except that the value function is estimated via MC instead of value iteration that relies on model parameters.

Algorithmically, one initializes an arbitrary policy π and an arbitrary function Q on $\mathcal{X} \times \mathcal{A}$, and keeps looping over the following steps:

- Pick X_0, A_0 such that all state-action pairs are sampled infinitely often;
- Generate an episode starting from X_0, A_0 following π;
- Update Q using first-visit (or every-visit) MC based on this trajectory, and after each update of Q, also modify π to be greedy with respect to Q.

Convergence of MCES does not hold for arbitrary MDPs, but results for important classes of MDPs have been established. In particular, if an MDP satisfies the Optimal Policy Feed-Forward (OPFF) assumption, which means that under any optimal policy a state is never re-visited, then first-visit MCES converges to the optimal value function almost surely [5]. Such MDPs include a large class of environments such as all episodic environments with a timestep (or any monotonically changing values as part of the state), as well as MDPs for which X_{t+1} and R_{t+1} are deterministic given (X_t, A_t).

2.2.2 On-Policy MC Control for Soft Policies

One way to eliminate the exploring starts assumption is to use soft policies in GPI, i.e. policies which is fully supported on $\mathcal{A}(x)$ for all states x. In particular, we say that a policy π is ϵ-soft if $\pi(a \mid x) \geq \epsilon/|\mathcal{A}(x)|$ for all state-action pairs (x, a). Among ϵ-soft policies are ϵ-greedy policies which, at each state x, assign probability $1 - \epsilon + \epsilon/|\mathcal{A}(x)|$ to some greedy action $a = \arg\max Q(x, \cdot)$ and probability $\epsilon/|\mathcal{A}(x)|$ to all the other actions $\mathcal{A}(x) \setminus \{a\}$. The following proposition shows that ϵ-greedy policies result in improvements.

Proposition 2.1. *Let π be ϵ-soft and π' be ϵ-greedy with respect to Q^π. Then $\pi \leq \pi'$, and the equality holds only when both π and π' are optimal among ϵ-soft policies.*

Proof. To show $\pi \leq \pi'$, we appeal to the policy improvement theorem, whose assumption is indeed satisfied since for all state x,

$$\begin{aligned} Q^\pi(x, \pi'(x)) = \sum_a \pi'(a \mid x) Q^\pi(x, a) &= \frac{\epsilon}{|\mathcal{A}(x)|} \sum_a Q^\pi(x, a) + (1 - \epsilon) \max_a Q^\pi(x, a) \\ &\geq \frac{\epsilon}{|\mathcal{A}(x)|} \sum_a Q^\pi(x, a) + (1 - \epsilon) \sum_a \frac{\pi(a \mid x) - \frac{\epsilon}{|\mathcal{A}(x)|}}{1 - \epsilon} Q^\pi(x, a) \\ &= \sum_a \pi(a \mid x) Q^\pi(x, a) = V^\pi(x), \end{aligned}$$

where the inequality follows from the assumption that π is ϵ-soft, so that a weighted average is no larger than the maximum.

Next we show that that if π' does not improve upon π then both are optimal among ϵ-soft policies. The idea is to move the "ϵ-softness" from policy into the environment. Specifically, modify the environment so that given any state-action pair (x, a), the new environment transitions like the old environment $\mathcal{P}_0(\cdot, \cdot \mid x, a)$ with probaiblity $1 - \epsilon$, but with probability ϵ, the new environment re-picks an action $a' \in \mathcal{A}(x)$ with uniform probability and then transitions according to $\mathcal{P}_0(\cdot, \cdot \mid x, a')$. Then the optimal state-value function in the new environment corresponds to the value function in the old environment that is optimal across ϵ-soft policies. It then suffices to note that the equation

$$V^\pi(x) = \frac{\epsilon}{|\mathcal{A}(x)|} \sum_a Q^\pi(x, a) + (1 - \epsilon) \max_a Q^\pi(x, a)$$

given by the derivation above is in fact the Bellman optimality equation for $\tilde{V}^\star$, the value function in the modified environment, so by uniqueness we have $V^\pi = \tilde{V}^\star$. □

We now state the the on-policy MC control for ϵ-soft policies algorithmically. First, initialize an arbitrary ϵ-soft policy π and an arbitrary $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, and then keep looping over the following steps:

- Generate an episode following π;
- Update Q using first-visit MC based on this trajectory, and after each update of Q, also modify π to be ϵ-greedy with respect to Q.

Note that upon convergence (if it happens), we obtain $\tilde{V}^\star$ instead of $V^\star$. In other words, this method learns action values for a near-optimal policy that still explores instead of action values for an optimal policy. It is therefore natural to consider separating the behavior policy and the target policy, i.e. off-policy Monte Carlo.

2.3 Off-Policy Monte Carlo

In off-policy methods, we separate the behavior policy b for sampling and the target policy π to be learnt. Off-policy methods are more general (taking $b = \pi$ recovers the on-policy case) and more powerful. For example, they enables learning from data generated by some external source (e.g. human expert). However, they are often of greater variance and slower to converge.

In order to estimate values for π using episodes from b, we impose the assumption of coverage: for any state x we have that $\pi(a \mid x) > 0$ implies $b(a \mid x) > 0$ for all action a. In control, π is often deterministic (e.g. greedy) while b remains stochastic and more exploratory (e.g. ϵ-greedy). In what follows we first discuss off-policy MC prediction and then move on to off-policy MC control.

2.3.1 Prediction

Almost all off-policy methods utilizes *importance sampling*, which is a general technique for estimating expected values under one distribution given samples from another. We apply this concept to behavior vs target policy.

Let $\tau = X_t, A_t, X_{t+1}, A_{t+1}, \cdots, X_T$ be a trajectory starting at X_t. The probability that π samples this trajectory is

$$\mathbb{P}_\pi(\tau \mid X_t) = \prod_{k=t}^{T-1} \pi(A_k \mid X_k)\mathcal{P}(X_{k+1} \mid X_k, A_k)$$

while the probability that b samples this trajectory is

$$\mathbb{P}_b(\tau \mid X_t) = \prod_{k=t}^{T-1} b(A_k \mid X_k)\mathcal{P}(X_{k+1} \mid X_k, A_k).$$

The *importance-sampling ratio* of this trajectory given X_t is defined by

$$\rho_{t:T-1} := \frac{\mathbb{P}_\pi(\tau \mid X_t)}{\mathbb{P}_b(\tau \mid X_t)} = \prod_{k=t}^{T-1} \frac{\pi(A_k \mid X_k)}{b(A_k \mid X_k)},$$

if $\mathbb{P}_\pi(\tau \mid X_t) > 0$ and zero otherwise. This is well-defined by the coverage assumption. As the notation suggests, we write $\rho_k := \pi(A_k \mid X_k)/b(A_k \mid X_k)$ and $\rho_{t_1:t_2} := \prod_{k=t_1}^{t_2} \rho_k$. Similarly, the ratio for τ given X_t, A_t is given by

$$\frac{\mathbb{P}_\pi(\tau \mid X_t, A_t)}{\mathbb{P}_b(\tau \mid X_t, A_t)} = \rho_{t+1:T-1}.$$

The importance sampling ratio allows us to rewrite the value functions for π

$$V^\pi(x) = \mathbb{E}_\pi[G_t \mid X_t = x] = \mathbb{E}_b[\rho_{t:T-1} G_t \mid X_t = x] \tag{11}$$

$$Q^\pi(x) = \mathbb{E}_\pi[G_t \mid X_t = x, A_t = a] = \mathbb{E}_b[\rho_{t+1:T-1} G_t \mid X_t = x, A_t = a] \tag{12}$$

as an expectation with respect to b. Note that the importance-sampling ratio involves only the two policies and not the model parameters, so that MC can be applied.

For notational convenience we label time steps across episode boundaries: if an episode starts at time $t \geq 0$, let $T(t)$ denote the termination time of this episode and start the next episode at time $T(t) + 1$. In addition, for each state x define $\mathcal{T}(x)$ to be the set of time steps t at which $X_t = x$ in the every-visit case; in the first-visit case define this to only include time steps that are first visits to x in each episode.

We consider two formulas to approximate the above expectation. Here we discuss the case for V^π, but the case for Q^π is analogous. The *ordinary importance sampling* (OIS) uses a simple average

$$V_{OIS}(x) := \frac{\sum_{t \in \mathcal{T}(x)} \rho_{t:T(t)-1} G_t}{|\mathcal{T}(x)|},$$

while the *weighted importance sampling* (WIS) is defined by

$$V_{WIS}(x) := \frac{\sum_{t \in \mathcal{T}(x)} \rho_{t:T(t)-1} G_t}{\sum_{t \in \mathcal{T}(x)} \rho_{t:T(t)-1}} = \frac{V_{OIS}(x)}{\sum_{t \in \mathcal{T}(x)} \rho_{t:T(t)-1} / |\mathcal{T}(x)|}.$$

We compare the two methods in terms of biases and variances:

- First-visit OIS gives an unbiased estimate for $V^\pi(x)$ by Equation (11), but its variance can be unbounded if the variance of $\rho_{t:T(t)-1}$ is unbounded (e.g. take the returns to be some fixed constant; see example in [3]), which makes convergence slower.

- First-visit WIS is biased but asymptotically unbiased, and (assuming bounded rewards) its variance is bounded when the importance sampling ratios have infinite variance, so this estimate is preferred in practice.

  To see this, note that $\frac{1}{|\mathcal{T}(x)|} \sum_{t \in \mathcal{T}(x)} \rho_{t:T(t)-1}$ converges to $\mathbb{E}_b[\rho_{t:T(t)-1} \mid X_t = x] = 1$ almost surely. Hence $V_{WIS}(x)$ converges to $V^\pi(x)$ almost surely as $|\mathcal{T}(x)| \to \infty$, so its expectation converges to $V^\pi(x)$. Now assume the maximum return is bounded (e.g. $\gamma < 1$ or the length of episode is bounded, together with the assumption that rewards are bounded), then $V_{WIS}(x)$ is bounded as it is a convex combination of returns, so the variance of $V_{WIS}(x)$ is bounded.

- The every-visit version of both estimates are biased but asymptotically unbiased. This is preferred in practice because there is no need to keep track of the first visits in each episode.

In fact, OIS and WIS are the least-squares solutions to the empirical objectives

$$V_{OIS}(x) = \arg\min_v \frac{1}{n}\sum_{k=1}^{n}(\rho_k G_k - v)^2,\ \ V_{WIS}(x) = \arg\min_v \frac{1}{n}\sum_{k=1}^{n}\rho_k(G_k - v)^2.$$

This observation helps extend WIS for linear function approximation [6].

When implementing these algorithms, it is important to use incremental update rules for better performance. The structure of ρ allows for natural incremental update, while the estimates should be updated via

$$V_{n+1} := V_n + \frac{W_n}{C_n}[G_n - V_n]\ \ (n \geq 1)$$

where G_n are returns starting from some state, $C_{n+1} := C_n + W_{n+1}$, $C_0 = 0$, $W_n := \rho_{t_n:T(t_n)-1}$, and V_1 arbitrary. These update rules give

$$V_n = \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}\ \ (n \geq 2).$$

The update rules for the WIS estimate of Q^π is analogous.

2.3.2 Control

A version of off-policy MC control can be found in [3], which involves off-policy prediction described above while modifying the target policy as greedy to the estimated action-value function. In this GPI, we have the freedom to change b from one episode to another (while maintaining the coverage of π), but it is often chosen to be the ϵ-greedy policy for the current value function to maintain some degree of exploration, otherwise it would be very slow as the algorithm keeps existing the inner loop (i.e. move on to the next episode) when the updated π differs from b (making the importance sampling ratio 0 and hence no need to update further). An implementation of the racetrack exercise in [3] can be found here.

References

[1] C. Szepesvari, *Algorithms for Reinforcement Learning.* 2019.

[2] A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun, *Reinforcement Learning: Theory and Algorithms.* 2022.

[3] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction.* 2018.

[4] S. P. Singh and R. S. Sutton, "Reinforcement learning with replacing eligibility traces," *Machine learning*, vol. 22, pp. 123–158, 1996.

[5] C. Wang, S. Yuan, K. Shao, and K. Ross, "On the convergence of the monte carlo exploring starts algorithm for reinforcement learning," *arXiv preprint arXiv:2002.03585*, 2020.

[6] A. R. Mahmood, H. P. Van Hasselt, and R. S. Sutton, "Weighted importance sampling for off-policy learning with linear function approximation," *Advances in neural information processing systems*, vol. 27, 2014.