# Reinforcement Learning

Siddhant Chaudhary

**Abstract**

These are my supplementary notes for a course on Reinforcement Learning which I took at CMI. The reference book used for the course was *Reinforcement Learning: An Introduction* by Richard S. Sutton and Andrew G. Barto.

## Contents



# 1. Markov Decision Processes

## 1.1 Defining MDPs

**1.1.1 Formal Definition.** A *Markov Decision Process* is a tuple $\langle S, A, P, R \rangle$, where the symbols mean the following.

- $S$ is a finite set of states.
- $A$ is a finite set of actions.
- $P$ is a transition function

$$P : S \times A \to \Delta(S)$$

Here $\Delta(S)$ is the set of all probability distributions on $S$. We use the notation $\mathbf{P}_{s,s'}(a)$ to denote the probability of reaching state $s'$ from state $s$ under the action $a$.

- $R$ is a reward function

$$R : S \times A \to \mathbf{R}$$

We assume that $R(s, a) \in [R_{\min}, R_{\max}]$ for all $s, a \in S \times A$.

**1.1.2 Policies and Value Functions.** A *policy* is a map $\pi : S \to A$. Note that by our assumptions on the finiteness of $A$ and $S$, there are only finitely many policies.

We now define the notion of an *optimal policy* for an MDP. To do that, we first define the *value* of a state. Let $\gamma \in [0, 1)$ be any number, which we'll call the *discount factor*. Under a policy $\pi$, we define the *value* of a state $s \in S$ by

$$V^\pi(s) := \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t)) \;\middle|\; s_0 = s\right]$$

We use the notation $V^\pi$ to denote the vector whose components are $V^\pi(s)$ for $s \in S$.

**1.1.3 Matrix equation for the value vector.** In this section, we will derive a closed form solution for the value vector $V^\pi$.

First, we have the following by the linearity of expectation.

$$\begin{aligned}
V^\pi(s) &= \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t)) \;\middle|\; s_0 = s\right] \\
&= \mathbf{E}\left[R(s_0, \pi(s_0)) + \sum_{t=1}^{\infty} \gamma^t R(s_t, \pi(s_t)) \;\middle|\; s_0 = s\right] \\
&= R(s, \pi(s)) + \mathbf{E}\left[\sum_{t=1}^{\infty} \gamma^t R(s_t, \pi(s_t)) \;\middle|\; s_0 = s\right] \\
&= R(s, \pi(s)) + \gamma \mathbf{E}_{s_1 \sim P(s,\pi(s))}\left[\sum_{t=0}^{\infty} \gamma^t R(s_{t+1}, \pi(s_{t+1}))\right] \\
&= R(s, \pi(s)) + \gamma\left(\sum_{s' \in S} \mathbf{P}_{s,s'}(\pi(s)) \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_{t+1}, \pi(s_{t+1})) \;\middle|\; s_1 = s'\right]\right) \\
&= R(s, \pi(s)) + \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(\pi(s)) V^\pi(s')
\end{aligned}$$

In other words, if we treat $V^\pi(s)$ as an unknown, then we obtain $S$ linear equations in $S$ unknowns, which we can solve easily. The above equation may be written in the following matrix form.

$$V^\pi = R^\pi + \gamma P^\pi V^\pi \tag{1}$$

Here, $V^\pi \in \mathbf{R}^{|S|}$ is the vector whose components are the values under policy $\pi$ of each state; $R^\pi \in \mathbf{R}^{|S|}$ is the reward vector, whose $s$th component is $R(s, \pi(s))$ and $P^\pi$ is an $|S| \times |S|$ matrix, whose $s$th row for $s \in S$ is the probability vector whose entries are $\mathbf{P}_{s,s'}(\pi(s))$ for $s' \in S$.

From (1), we can see that

$$(I - \gamma P^\pi) V^\pi = R^\pi$$

where $I$ is the $|S| \times |S|$ identity matrix. Clearly, if $(I - \gamma P^\pi)$ was invertible, then we would have

$$V^\pi = (I - \gamma P^\pi)^{-1}(R^\pi) \tag{2}$$

We will now show that this is indeed the case.

**Proposition 1.1.** *The matrix $I - \gamma P^\pi$ where $\gamma \in (0,1)$ is invertible.*

*Proof.* Suppose $x \in \text{Ker}(I - \gamma P^\pi)$, i.e

$$(I - \gamma P^\pi)x = 0$$

This would imply that

$$x = \gamma P^\pi x$$

which would imply that $1/\gamma$ is an eigenvalue of $p^\pi$, if $x \neq 0$. Note that $1/\gamma > 1$. But because $P^\pi$ is row-stochastic, all it's eigenvalues have to be $\leq 1$. So, it must be the case that $x = 0$, i.e $I - \gamma P^\pi$ is invertible. ■

**1.1.4 Inverse of $I - \gamma P^\pi$.** Consider the matrix

$$\sum_{t=0}^{\infty} \gamma^t (P^\pi)^t$$

This series converges, and infact this matrix is equal to $(I - \gamma P^\pi)$. We denote this matrix by $D^\pi$. Intuitively, the entries of the matrix $D^\pi$ reflect the *number of times* we visit a state if we start from a given state under the policy $\pi$. From (2), we see that

$$V^\pi = D^\pi R^\pi$$

**1.1.5 The Bellman Backup Operator.** In this section, we will define the so called *Bellman Backup Operator* and prove some of it's properties.

For a policy $\pi$, we define a map $B_\pi : \mathbf{R}^{|S|} \to \mathbf{R}^{|S|}$ by the following.

$$B_\pi(v) = R^\pi + \gamma P^\pi v \qquad , \qquad v \in \mathbf{R}^{|S|}$$

As usual, $\gamma \in (0,1)$ is some number. We will now prove some properties of this map, in the context of an MDP.

**Proposition 1.2 (Monotonicity).** *Let $u, v \in \mathbf{R}^{|S|}$ and let $\leq$ be the partial order in $\mathbf{R}^{|S|}$ which compares vectors component wise. Then,*

$$u \leq v \implies B_\pi(u) \leq B_\pi(v)$$

*Proof.* We have the following.

$$\begin{aligned} u \leq v &\implies P^\pi u \leq P^\pi v && \text{(all entries of } P^\pi \text{ are non-negative)} \\ &\implies \gamma P^\pi u \leq \gamma P^\pi v \\ &\implies R^\pi + \gamma P^\pi u \leq R^\pi + \gamma P^\pi v \\ &\implies B_\pi(u) \leq B_\pi(v) \end{aligned}$$

■

**Proposition 1.3 (Contraction).** *Let $u, v \in \mathbf{R}^{|S|}$. Then,*

$$||B_\pi(u) - B_\pi(v)||_\infty \leq \gamma \, ||u - v||_\infty \tag{3}$$

*Proof.* We will show that for each $s \in S$, we have

$$|B_\pi(u)(s) - B_\pi(v)(s)| \leq \gamma \, ||u - v||_\infty$$

Here, $B_\pi(u)(s)$ is the $s$th coordinate of the vector $B_\pi(u)$ (and similarly for $B_\pi(v)$). We have the following.

$$\begin{aligned}
|B_\pi(u)(s) - B_\pi(v)(s)| &= \left| \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(\pi(s)) u(s') - \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(\pi(s)) v(s') \right| \\
&= \gamma \left| \sum_{s' \in S} \mathbf{P}_{s,s'}(\pi(s)) [u(s') - v(s')] \right| \\
&\leq \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(\pi(s)) \left| u(s') - v(s') \right| \\
&\leq \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(\pi(s)) \, ||u - v||_\infty \\
&= \gamma \, ||u - v||_\infty
\end{aligned}$$

Above, we used the fact that $\sum_{s' \in S} \mathbf{P}_{s,s'}(\pi(s)) = 1$. This proves the claim. ■

Since $\mathbf{R}^{|S|}$ is a complete space, we can use the contraction property of $B_\pi$ and conclude by the **Contraction Mapping Theorem** that $V^\pi$ is the unique fixed point of $B_\pi$, and moreover for any vector $v \in \mathbf{R}^{|S|}$ we have

$$V^\pi = \lim_{n \to \infty} B_\pi^n(v) \tag{4}$$

**1.1.6 $Q$-values.** In this section we will introduce another notation which will be a bit convenient.

Let $s \in S$ be any state, $a \in A$ be any action and $v \in \mathbf{R}^{|S|}$ be any vector. We define the *one step backup of action $a$ at state $s$ with respect to vector $v$* by

$$Q^v(s, a) := R(s, a) + \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(a) v(s')$$

If $\pi$ is a policy, we also use the notation $Q^\pi(s, \pi(s))$ to refer to the quantity $Q^{V^\pi}(s, \pi(s))$ (recall that the vector $V^\pi$ is the unique fixed point of $B_\pi$).

**1.1.7 Existence of the best policy.** We will now come to the one of the most important results related to what we've seen thus far, which is the existence of the *best policy* for an MDP.

As usual, let $(S, A, P, R)$ be an MDP. On the space $\mathbf{R}^{|S|}$, consider the partial order $\leq$, which compares two vectors componentwise. We will now show that

$$\pi^* = \text{argmin}_\pi V^\pi$$

exists, where the minimum is taken with respect to the partial order $\leq$ we mentioned above. In simple words, for any MDP, there is a policy $\pi$ such that the value vector $V^\pi$ associated to that policy has the property that all of it's coordinates are greater than or equal to the coordinates of any other value vector for any other policy. We will show this as a sequence of results.

**Theorem 1.4.** *Let $(S, A, P, R)$ be an MDP, and let $\pi_1$, $\pi_2$ be any two policies. Then, there is a policy $\pi$ such that*

$$V_{\max} \leq V^{\pi}$$

*where the vector $V_{\max}$ is defined as*

$$V_{\max}(s) = \max(V^{\pi_1}(s), V^{\pi_2}(s))$$

*for all $s \in S$.*

*Proof.* Consider the following simple policy.

$$\pi(s) := \begin{cases} \pi_1(s) & , \quad \text{if } V^{\pi_1}(s) \geq V^{\pi_2}(s) \\ \pi_2(s) & , \quad \text{otherwise} \end{cases}$$

We will now show that the vector $V^{\pi}$ does the job.

To show this, first note that $V^{\pi_1} \leq V_{\max}$ and $V^{\pi_2} \leq V_{\max}$. Using this, we will show that

$$V_{\max} \leq B_{\pi}(V_{\max}) \tag{5}$$

Note that to show the above, it is enough to show that $V^{\pi_1} \leq B_{\pi}(V_{\max})$ and $V^{\pi_2} \leq B_{\pi}(V_{\max})$. Let's show the first of these two inequalities, as the second one has the exact same proof. So let $s \in S$. First, suppose $V^{\pi_1}(s) \geq V^{\pi_2}(s)$, and in that case we will have that $\pi(s) = \pi_1(s)$. In that case, we have

$$\begin{aligned} V^{\pi_1}(s) = B_{\pi_1}(V^{\pi_1})(s) &= R(s, \pi_1(s)) + \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(\pi_1(s)) V^{\pi_1}(s') \\ &= R(s, \pi(s)) + \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(\pi(s)) V^{\pi_1}(s') \\ &\leq R(s, \pi(s)) + \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(\pi(s)) V_{\max}(s') \\ &= B_{\pi}(V_{\max})(s) \end{aligned}$$

In the other case, suppose that $V^{\pi_1}(s) < V^{\pi_2}(s)$, which will mean that $\pi(s) = \pi_2(s)$. In that case, we have the following.

$$\begin{aligned} V^{\pi_1}(s) < V^{\pi_2}(s) = B_{\pi_2}(V^{\pi_2})(s) &= R(s, \pi_2(s)) + \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(\pi_2(s)) V^{\pi_2}(s') \\ &= R(s, \pi(s)) + \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(\pi(s)) V^{\pi_2}(s') \\ &\leq R(s, \pi(s)) + \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(\pi(s)) V_{\max}(s') \\ &= B_{\pi}(V_{\max})(s) \end{aligned}$$

So we've shown that $V^{\pi_1} \leq B_{\pi}(V_{\max})$ and $V^{\pi_2} \leq B_{\pi}(V_{\max})$ and hence (5) follows.

Now, by the monotonicity (**Proposition** 1.2) of $B_{\pi}$, we see that the sequence $\{B_{\pi}^n(V_{\max})\}_{n \in \mathbf{N}}$ is a non-decreasing sequence. So, this implies that

$$V_{\max} \leq \lim_{n \to \infty} B_{\pi}^n(V_{\max}) = V^{\pi}$$

and this completes the proof of the claim. ■

**Theorem 1.5.** *Let $(S, A, P, R)$ be an MDP. Then, there is a policy $\pi^*$ such that for all policies $\pi$*

$$V^{\pi} \leq V^{\pi^*}$$

*Proof.* Since there are only finitely many policies, we can enumerate them; let $\{\pi_1, \pi_2, ..., \pi_K\}$ be all the policies. The claim follows by a simple application of **Theorem** 1.4. More precisely, define $\pi'_1 := \pi_1$, and for each $2 \leq k \leq K$ define $\pi'_k$ to be the policy that is $\geq \pi_k$ and $\pi'_{k-1}$ (possible because of **Theorem** 1.4). Then, the policy $\pi'_K$ will be the desired policy. ■

**1.1.8 Bellman Optimality Conditions.** From what we've seen till now, the optimal value vector $V^*$ satisfies the following equations.

$$V^*(s) = \max_{a \in A} \left\{ R(s,a) + \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(a) V^*(s') \right\} = \max_{a \in A} Q^{V^*}(s,a)$$

This is true because, if there is some state $s$ such that $V^*(s)$ is not equal to the given maximum, then we can find a better policy by using the action which attains the maximum, and thereby we will obtain a value vector which is strictly larger than $V^*$, which will be a contradiction. These equations are known as the *Bellman Optimality Conditions.*

Similarly, an optimal policy $\pi^*$ satisfies the following.

$$\pi^*(s) \in \text{argmax}_{a \in A} \left\{ R(s,a) + \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(a) V^*(s') \right\} = \text{argmax}_{a \in A} Q^{V^*}(s,a)$$

**1.1.9 Bellman Optimality Operator and the Value Iteration Algorithm.** Like like the Bellman Backup Operator, we will now define the *Bellman Optimality Operator*; it is a map $B : \mathbf{R}^{|S|} \to \mathbf{R}^{|S|}$ defined by the following.

$$B[V](s) := \max_{a \in A} \left\{ R(s,a) + \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(a) V(s') \right\} = \max_{a \in A} Q^{V}(s,a)$$

As before, $B$ also satisfies the properties of monotonicity and contraction.

**Proposition 1.6.** *Let $B$ be the Bellman Optimality Operator as defined above. Then, $B$ satisfies the properties of monotonicity and contraction (w.r.t the $||\cdot||_\infty$ norm) as mentioned in **Proposition** 1.2 and **Proposition** 1.3.*

*Proof.* These are not hard to prove, and can be done similar to the proofs in **Proposition** 1.2 and **Proposition** 1.3. ■

Just like before, we can apply the contraction property to see that for any $v \in \mathbf{R}^{|S|}$, we have

$$V^* = \lim_{k \to \infty} B^k(v)$$

The above fact leads to the following simple algorithm algorithm to compute $V^*$, which is called the *Value Iteration Algorithm.*

**Algorithm 1** Value Iteration (VI) Algorithm

$V_0 \leftarrow 0$.
$i \leftarrow 0$.
**while** $V_i \neq V_{i-1}$ **do**
  For each $s \in S$, $V_{i+1}(s) \leftarrow \max_{a \in S} Q^{V_i}(s, a)$
  $i \leftarrow i + 1$
**end while**
**return** $V_i$.

**1.1.10 Stopping early in the VI algorithm.** In practice, instead of stopping when $V_i = V_{i-1}$, the usual stopping criterion is

$$||V_i - V_{i-1}||_\infty \leq \epsilon$$

where $\epsilon > 0$ is some tolerance level. We now bound the distance of $V_i$ from $V^*$ if such a stopping criterion is used.

**Proposition 1.7.** *Suppose the VI algorithm stops at a point when* $||V_i - V_{i-1}||_\infty \leq \epsilon$. *Then*

$$||V_i - V^*||_\infty \leq \frac{\epsilon\gamma}{1-\gamma}$$

*where* $V^*$ *is the optimal value function.*

*Proof.* Note that $V_i = B[V_{i-1}]$, where $B$ is the Bellman Optimality Operator. Moreover, we know that $B[V^*] = V^*$. So, by the contraction property, we see that

$$||V_i - V^*||_\infty = ||B[V_{i-1}] - B[V^*]||_\infty \leq \gamma\, ||V_{i-1} - V^*||_\infty$$

Also, by the triangle inequality we have that

$$\begin{aligned} \gamma\, ||V_{i-1} - V^*||_\infty &\leq \gamma\, ||V_i - V^*||_\infty + \gamma\, ||V_i - V_{i-1}||_\infty \\ &\leq \gamma\, ||V_i - V^*||_\infty + \epsilon\gamma \end{aligned}$$

So, combining the last two inequalities, we see that

$$||V_i - V^*||_\infty \leq \gamma\, ||V_i - V^*||_\infty + \epsilon\gamma$$

and from here the claim follows. ■

**Definition 1.1 (Greedy policy with respect to a vector.).** Let $V \in \mathbf{R}^{|S|}$ be any vector. A *greedy policy* with respect to $V$ is a policy $\pi^V$ such that

$$\pi^V(s) \in \text{argmax}_{a \in A} Q^V(s, a)$$

for all $s \in S$.

We will now see how good the greedy policy of the output vector of VI algorithm is, if we use the stopping criterion with tolerance $\epsilon$.

**Proposition 1.8.** *Suppose the VI algorithm stops at the vector* $\tilde{V}$, *and suppose* $\left|\left|\tilde{V} - V^*\right|\right|_\infty \leq \delta$. *Let* $\tilde{\pi}$ *be a greedy policy with respect to* $\tilde{V}$. *Then,*

$$\left|\left|V^{\tilde{\pi}} - V^*\right|\right|_\infty \leq \frac{2\delta\gamma}{1-\gamma}$$

*Proof.* First, we claim that

$$B[\tilde{V}] = B_{\tilde{\pi}}[\tilde{V}]$$

But this is easy to see by the definition of the Bellman Optimality Operator and since $\tilde{\pi}$ is a greedy policy with respect to $\tilde{V}$. Now, using this fact we have the following.

$$\begin{aligned}
\left|\left|V^{\tilde{\pi}} - V^*\right|\right|_\infty &\leq \left|\left|V^{\tilde{\pi}} - B[\tilde{V}] + B[\tilde{V}] - V^*\right|\right|_\infty \\
&= \left|\left|B_{\tilde{\pi}}[V^{\tilde{\pi}}] - B_{\tilde{\pi}}[\tilde{V}] + B[\tilde{V}] - V^*\right|\right|_\infty && (V^{\tilde{\pi}} \text{ is a fixed point of } B_{\tilde{\pi}}) \\
&\leq \left|\left|B_{\tilde{\pi}}[V^{\tilde{\pi}}] - B_{\tilde{\pi}}[\tilde{V}]\right|\right|_\infty + \left|\left|B[\tilde{V}] - V^*\right|\right|_\infty \\
&\leq \gamma \left|\left|V^{\tilde{\pi}} - \tilde{V}\right|\right|_\infty + \left|\left|B[\tilde{V}] - B[V^*]\right|\right|_\infty && (V^* \text{ is a fixed point of } B) \\
&\leq \gamma \left|\left|V^{\tilde{\pi}} - \tilde{V}\right|\right|_\infty + \gamma \left|\left|\tilde{V} - V^*\right|\right|_\infty \\
&\leq \gamma \left|\left|V^{\tilde{\pi}} - \tilde{V}\right|\right|_\infty + \gamma\delta \\
&= \gamma \left|\left|V^{\tilde{\pi}} - V^* + V^* - \tilde{V}\right|\right|_\infty + \gamma\delta \\
&\leq \gamma \left|\left|V^{\tilde{\pi}} - V^*\right|\right|_\infty + \gamma \left|\left|V^* - \tilde{V}\right|\right|_\infty + \gamma\delta \\
&\leq \gamma \left|\left|V^{\tilde{\pi}} - V^*\right|\right| + 2\gamma\delta
\end{aligned}$$

From this, the claim follows. ■

**1.1.11 Stochastic Policies.** A *stochastic policy* $\pi$ is a policy which assigns a probability distribution over the actions to each state; formally, for each $a \in A$ and $s \in S$, the probability of taking action $a$ in state $S$ under the policy $\pi$ has probability $\pi(a|s)$, and hence we must have

$$\sum_{a \in A} \pi(a|s) = 1$$

For a stochastic policy $\pi$, we define the immediate reward starting from a state by

$$\sum_{a \in A} \pi(a|s) R(s, a)$$

So in this setting, the equation for the value vector $V^\pi$ becomes the following.

$$V^\pi(s) = \sum_{a \in A} \pi(a|s) \left[ R(s, a) + \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(a) \cdot V^\pi(s') \right]$$

Again, we can write the above equation in matrix form.

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

Let's see what the vector $R^\pi$ and the matrix $P^\pi$ are in this case. The $s$th entry of the reward vector $R^\pi$ is simply

$$\sum_{a \in A} \pi(a|s) R(s, a)$$

Similarly, the $s$th row of the matrix $P^\pi$ will be just be the weighted sum of the rows indexed by the state-action pairs $(s, a)_{a \in A}$, where the weights will be $\pi(a|s)_{a \in A}$.

**1.1.12 MDPs as Linear Programs.** MDPs can also be written as LPs. In this section, we will see how to do this.

Since we want to solve for $V^*$, the variables for our LP will be $V^*(s_1)$, ..., $V^*(s_n)$, where $s_1, ..., s_n$ are the states. The constraints of the LP will be a consequence of the Bellman Optimality conditions; in particular, for each $s \in S$ and $a \in A$, we will have a constraint that specifies

$$V^*(s) \geq Q^*(s,a) = R(s,a) + \sum_{s' \in S} \mathbf{P}_{s,s'}(a) V^*(s')$$

So, there will be $nm$ constraints, if $|S| = n$ and $|A| = m$. Finally, our objective will be to minimize the quantity

$$\sum_{s \in S} V^*(s)$$