Reinforcement Learning **Spring Semester, 2018/9**

Lecture 4: March 20, 2019

Lecturer: Yishay Mansour *Scribe: ym*

DISCLAIMER: Based on `Learning and Planning in Dynamical Systems` by Shie Mannor©, all rights reserved.

This lecture covers the basic theory and main solution methods for stationary MDPs over an infinite horizon, with the discounted return criterion. In this case, stationary policies are optimal.

The discounted return problem is the most "well behaved" among all infinite horizon problems (such as average return and stochastic shortest path), and the theory of it is relatively simple, both in the planning and the learning contexts. For that reason, as well as its usefulness, we will consider here the discounted problem and its solution in some detail.

4.1 Problem Statement

We consider a stationary (time-invariant) MDP, with a finite state space S, finite action set A, and transition kernel $P = (P(s'|s,a))$ over the infinite time horizon $\mathbb{T} = \{0, 1, 2, \ldots\}$.

Our goal is to maximize the expected discounted return, which is defined for each control policy π and initial state $s_0 = s$ as follows:

$$\begin{aligned} J_\gamma^\pi(s) &= E^\pi(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|s_0 = s) \\ &\equiv E^{\pi,s}(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)) \end{aligned}$$

Here,

- $r : S \times A \to \mathbb{R}$ is the (running, or instantaneous) expected reward function, i.e., $r(s,a) = E[R|s,a]$.
- $\gamma \in (0,1)$ is the discount factor.

We observe that $\gamma < 1$ ensures convergence of the infinite sum (since the rewards $r(s_t, a_t)$ are uniformly bounded). With $\gamma = 1$ we obtain the total return criterion, which is harder to handle due to possible divergence of the sum.

Let $J^*_\gamma(s)$ denote the maximal value of the discounted return, over all (possibly history dependent and randomized) control policies, i.e.,

$$J^*_\gamma(s) = \sup_{\pi \in \Pi_{HS}} J^\pi_\gamma(s).$$

Our goal is to find an optimal control policy π^* that attains that maximum (for all initial states), and compute the numeric value of the optimal return $J^*_\gamma(s)$. As we shall see, for this problem there always exists an optimal policy which is a (deterministic) stationary policy.

Note: As usual, the discounted performance criterion can be defined in terms of cost:

$$J^\pi_\gamma(s) = E^{\pi,s}(\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t))$$

where $c(s, a)$ is the running cost function. Our goal is then to minimize the discounted cost $J^\pi_\gamma(s)$.

4.2 The Fixed-Policy Value Function

We start the analysis by defining and computing the value function for a fixed stationary policy. This intermediate step is required for later analysis of our optimization problem, and also serves as a gentle introduction to the value iteration approach.

For a stationary policy $\pi : S \to A$, we define the value function $V^\pi(s),\ s \in S$ simply as the corresponding discounted return:

$$V^\pi(s) \stackrel{\Delta}{=} E^{\pi,s}\left(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right) = J^\pi_\gamma(s), \quad s \in S$$

Lemma 4.1. *V^π satisfies the following set of $|S|$ linear equations:*

$$V^\pi(s) = r(s, \pi(s)) + \gamma \sum\nolimits_{s' \in S} p(s'|s, \pi(s)) V^\pi(s'), \quad s \in S. \tag{4.1}$$

Proof. We first note that

$$\begin{aligned} V^\pi(s) &\triangleq E^\pi(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|s_0 = s) \\ &= E^\pi(\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t)|s_1 = s), \end{aligned}$$

since both the model and the policy are stationary. Now,

$$\begin{aligned} V^\pi(s) &= r(s, \pi(s)) + E^\pi(\sum_{t=1}^{\infty} \gamma^t r(s_t, \pi(s_t))|s_0 = s) \\ &= r(s, \pi(s)) + E^\pi\left[E^\pi\left(\sum_{t=1}^{\infty} \gamma^t r(s_t, \pi(s_t))|s_0 = s, s_1 = s' \right) \middle| s_0 = s \right] \\ &= r(s, \pi(s)) + \sum_{s' \in S} p(s'|s, \pi(s)) E^\pi(\sum_{t=1}^{\infty} \gamma^t r(s_t, \pi(s_t))|s_1 = s') \\ &= r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s'|s, \pi(s)) E^\pi(\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t)|s_1 = s') \\ &= r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s'|s, \pi(s)) V^\pi(s'). \end{aligned}$$

The first equality is by the smoothing theorem. The second equality follows since $s_0 = s$ and $a_t = \pi(s_t)$, the third equality follows similarly to the finite-horizon case (done in the previous lecture), the fourth is simple algebra, and the last by the observation above. □

We can write the linear equations in (4.1) in vector form as follows. Define the column vector $r^\pi = (r^\pi(s))_{s \in S}$ with components $r^\pi(s) = r(s, \pi(s))$, and the transition matrix P^π with components $P^\pi(s'|s) = p(s'|s, \pi(s))$. Finally, let V^π denote a column vector with components $V^\pi(s)$. Then (4.1) is equivalent to the linear equation set

$$V^\pi = r^\pi + \gamma P^\pi V^\pi \tag{4.2}$$

Lemma 4.2. *The set of linear equations* (4.1) *or* (4.2)*, with V^π as variables, has a unique solution V^π, which is given by*

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

Proof. We only need to show that the square matrix $I - \gamma P^\pi$ is non-singular. Let (λ_i) denote the eigenvalues of the matrix P^π. Since P^π is a stochastic matrix (row sums are 1), then $|\lambda_i| \leq 1$. Now, the eignevalues of $I - \gamma P^\pi$ are $(1 - \gamma\lambda_i)$, and satisfy $|1 - \gamma\lambda_i| \geq 1 - \gamma > 0$. □

Combining Lemma 4.1 and Lemma 4.2, we obtain

Proposition 4.1. *The fixed-policy value function $V^\pi = [V^\pi(s)]$ is the unique solution of equation (4.2), given by*

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

Proposition 4.1 provides a closed-form formula for computing V^π. However, for large systems, computing the inverse $(I - \gamma P^\pi)^{-1}$ may be hard. In that case, the following value iteration algorithm provides an alternative, iterative method for computing V^π.

Algorithm 4.1. ***Fixed-policy value iteration***

1. *Let $V_0 = (V_0(s))_{s \in S}$ be arbitrary.*
2. *For $n = 0, 1, 2, \ldots$, set*

$$V_{n+1}(s) = r(s, \pi(s)) + \gamma \sum\nolimits_{s' \in S} p(s'|s, \pi(s)) V_n(s') \,, \qquad s \in S$$

or, equivalently,

$$V_{n+1} = r^\pi + \gamma P^\pi V_n.$$

Proposition 4.2 (**Convergence of fixed-policy value iteration**). *We have $V_n \to V^\pi$ component-wise, that is,*

$$\lim_{n \to \infty} V_n(s) = V^\pi(s), \qquad s \in S.$$

Proof. Note first that

$$\begin{aligned} V_1(s) &= r(s, \pi(s)) + \gamma \sum\nolimits_{s' \in S} p(s'|s, \pi(s)) V_0(s') \\ &= E^\pi(r(s_0, a_0) + \gamma V_0(s_1) | s_0 = s). \end{aligned}$$

Continuing similarly, we obtain that

$$V_n(s) = E^\pi(\sum_{t=0}^{n-1} \gamma^t r(s_t, a_t) + \gamma^n V_0(s_n) | s_0 = s).$$

Note that $V_n(s)$ is the n-stage discounted return, with terminal reward $r_n(s_n) = V_0(s_n)$. Comparing with the definition of V^π, we can see that

$$V^\pi(s) - V_n(s) = E^\pi(\sum_{t=n}^{\infty} \gamma^t r(s_t, a_t) - \gamma^n V_0(s_n) | s_0 = s).$$

Denoting $R_{\max} = \max_{s,a}|r(s,a)|$, $\bar{V}_0 = \max_s |V_0(s)|$ we obtain

$$|V^\pi(s) - V_n(s)| \leq \gamma^n (\frac{R_{\max}}{1-\gamma} + \bar{V}_0)$$

which converges to 0 since $\gamma < 1$. □

Comments:

- The proof provides an explicit bound on $|V^\pi(s) - V_n(s)|$. It may be seen that the convergence is exponential, with rate $O(\gamma^n)$.
- Using vector notation, it may be seen that

$$V_n = r^\pi + P^\pi r^\pi + \ldots + (P^\pi)^{n-1} r^\pi + (P^\pi)^n V_0 = \sum_{t=0}^{n-1} (P^\pi)^t r^\pi + (P^\pi)^n V_0.$$

  Similarly, $V^\pi = \sum_{t=0}^{\infty} (P^\pi)^t r^\pi$.

In summary:

- Proposition 4.1 allows to compute V^π by solving a set of $|S|$ linear equations.
- Proposition 4.2 computes V^π by an infinite recursion, that converges exponentially fast.

4.3 Overview: The Main DP Algorithms

We now return to the optimal planning problem defined in Section 4.1. Recall that $J^*_\gamma(s) = \sup_{\pi \in \Pi_{HS}} J^\pi_\gamma(s)$ is the optimal discounted return. We further denote

$$V^*(s) \stackrel{\Delta}{=} J^*_\gamma(s), \quad s \in S,$$

and refer to V^* as the optimal value function. Depending on the context, we consider V^* either as a function $V^* : S \to \mathbb{R}$, or as a column vector $V^* = [V(s)]_{s\in S}$.

The following optimality equation provides an explicit characterization of the value function, and shows that an optimal stationary policy can easily be computed if the value function is known.

Theorem 4.1 (**Bellman's Optimality Equation**). *The following statements hold:*

1. *V^* is the unique solution of the following set of (nonlinear) equations:*

$$V(s) = \max_{a\in A} \left\{ r(s,a) + \gamma \sum\nolimits_{s'\in S} p(s'|s,a)V(s') \right\}, \quad s \in S. \tag{4.3}$$

2. *Any stationary policy π^* that satisfies*

$$\pi^*(s) \in \arg\max_{a\in A} \left\{ r(s,a) + \gamma \sum\nolimits_{s'\in S} p(s'|s,a)V(s') \right\},$$

is an optimal policy (for any initial state $s_0 \in S$).

The optimality equation (4.3) is non-linear, and generally requires iterative algorithms for its solution. The main iterative algorithms are **value iteration** and **policy iteration**.

Algorithm 4.2. ***Value iteration***

1. *Let $V_0 = (V_0(s))_{s\in S}$ be arbitrary.*
2. *For $n = 0, 1, 2, \ldots$, set*

$$V_{n+1}(s) = \max_{a\in A} \left\{ r(s,a) + \gamma \sum\nolimits_{s'\in S} p(s'|s,a)V_n(s') \right\}, \quad \forall s \in S$$

Theorem 4.2 (**Convergence of value iteration**). *We have $\lim_{n\to\infty} V_n = V^*$ (component-wise). The rate of convergence is exponential, at rate $O(\gamma^n)$.*

Proof. Using our previous results on value iteration for the finite-horizon problem, it follows that

$$V_n(s) = \max_{\pi} E^{\pi,s}\Big(\sum_{t=0}^{n-1} \gamma^t R_t + \gamma^n V_0(s_n)\Big).$$

Comparing to the optimal value function

$$V^*(s) = \max_{\pi} E^{\pi,s}\Big(\sum_{t=0}^{\infty} \gamma^t R_t\Big),$$

it may be seen that that

$$|V_n(s) - V^*(s)| \leq \gamma^n (\frac{R_{\max}}{1-\gamma} + ||V_0||_\infty).$$

As $\gamma < 1$, this implies that V_n converges to V_γ^* exponentially fast. □

The value iteration algorithm iterates over the value functions, with asymptotic convergence. The policy iteration algorithm iterates over stationary policies, with each new policy better than the previous one. This algorithm converges to the optimal policy in a finite number of steps.

Algorithm 4.3. ***Policy iteration***

1. *Initialization: choose some stationary policy* π_0.
2. *For* $k = 0, 1, \ldots$:
   (a) *Policy evaluation: compute* V^{π_k}. *% For example, use the explicit formula* $V^{\pi_k} = (I - \gamma P^{\pi_k})^{-1} r^{\pi_k}$.
   (b) *Policy Improvement: Compute* π_{k+1}, *a greedy policy with respect to* V^{π_k}:

$$\pi_{k+1}(s) \in \arg\max_{a \in A} \left\{ r(s,a) + \gamma \sum\nolimits_{s' \in S} p(s'|s,a) V^{\pi_k}(s') \right\}, \quad \forall s \in S.$$

   (c) *Stop if* $V^{\pi_{k+1}} = V^{\pi_k}$ *(or if* V^{π_k} *satisfied the optimality equation), else continue.*

Theorem 4.3 (**Convergence of policy iteration**). *The following statements hold:*

1. *Each policy* π_{k+1} *is improving over the previous one* π_k, *in the sense that* $V^{\pi_{k+1}} \geq V^{\pi_k}$ *(component-wise).*
2. $V^{\pi_{k+1}} = V^{\pi_k}$ *if and only if* π_k *is an optimal policy.*
3. *Consequently, since the number of stationary policies is finite,* π_k *converges to the optimal policy after a finite number of steps.*

An additional solution method for DP planning relies on a Linear Programming formulation of the problem. A Linear Program (LP) is simply an optimization problem with linear objective function and linear constraints. We will provide additional details later in this Lecture.

4.4 Contraction Operators

The basic proof methods of the DP results mentioned above rely on the concept of a *contraction operator*. We provide here the relevant mathematical background, and illustrate the contraction properties of some basic Dynamic Programming operators.

4.4.1 The contraction property

Recall that a norm $||\cdot||$ over $\mathbb{R}^n$ is a real-valued function $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}$ such that, for any pair of vectors $x, y \in \mathbb{R}^d$ and scalar a,

1. $||ax|| = |a| \cdot ||x||$,
2. $||x + y|| \leq ||x|| + ||y||$,
3. $||x|| = 0$ only if $x = 0$.

Common examples are the p-norm $||x||_p = (\sum_{i=1}^{d} |x_i|^p)^{1/p}$ for $p \geq 1$, and in particular the Euclidean norm ($p = 2$). Here we will mostly use the max-norm:

$$||x||_\infty = \max_{1 \leq i \leq d} |x_i|.$$

Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be a vector-valued function over $\mathbb{R}^d$ ($d \geq 1$). We equip $\mathbb{R}^d$ with some norm $||\cdot||$, and refer to T as an *operator* over $\mathbb{R}^d$. Thus, $T(v) \in \mathbb{R}^d$ for any $v \in \mathbb{R}^d$. We also denote $T^n(v) = T(T^{n-1}(v))$ for $n \geq 2$. For example, $T^2(v) = T(T(v))$.

Definition 4.1. *The operator T is called a contraction operator if there exists $\beta \in (0, 1)$ (the contraction coefficient) such that*

$$||T(v_1) - T(v_2)|| \leq \beta ||v_1 - v_2||,$$

for all $v_1, v_2 \in \mathbb{R}^d$

4.4.2 The Banach Fixed Point Theorem

The following celebrated result applies to contraction operators. While we quote the result for $\mathbb{R}^d$, we note that it applies in much greater generality to any Banach space (a complete normed space), or even to any complete metric space, with essentially the same proof.

Theorem 4.4 (**Banach's fixed point theorem**). *Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be a contraction operator. Then*

1. *The equation $T(v) = v$ has a unique solution $V^* \in \mathbb{R}^d$.*
2. *For any $v_0 \in \mathbb{R}^d$, $\lim_{n\to\infty} T^n(v_0) = V^*$. In fact, $||T^n(v_0) - V^*|| \leq O(\beta^n)$, where β is the contraction coefficient.*

Proof. Fix any v_0 and define $v_{n+1} = T(v_n)$. We will show that: (1) there exists a limit to the sequence, and (2) the limit is a fixed point of T.

Existence of a limit v^* of the sequence v_n
We show that the sequence of v_n is a cauchy sequence. We consider two elements v_n and v_{m+n} and bound the distance between them.

$$\begin{aligned}
\|v_{n+m} - v_n\| &= \|\sum_{k=0}^{m-1} v_{n+k+1} - v_{n+k}\| \\
&\leq \sum_{k=0}^{m-1} \|v_{n+k+1} - v_{n+k}\| \quad (\textit{according to the triangle inequality}) \\
&= \sum_{k=0}^{m-1} \|T^{n+k}v_1 - T^{n+k}v_0\| \\
&\leq \sum_{k=0}^{m-1} \beta^{n+k}\|v_1 - v_0\| \quad (\textit{contraction } n+k \textit{ times}) \\
&= \frac{\beta^n(1-\beta^m)}{1-\beta}\|v_1 - v_0\|
\end{aligned}$$

Since the coefficient decreases as n increases, for any $\epsilon > 0$ there exists $N > 0$ such that for all $n \geq N, \|\vec{v}_{n+m} - \vec{v}_n\| < \epsilon$. This implies that the sequence is a Cauchy sequence, and hence the sequence v_n has a limit. Let us call this limit v^*. Next we show that v^* is a fixed point of the operator T.

v^* is a fixed point
We need to show that $T(v^*) = v^*$, or equivalently $\|T(v^*) - v^*\| = 0$.

$$\begin{aligned}
0 &\leq \|T(v^*) - v^*\| \\
&\leq \|T(v^*) - v_n\| + \|v_n - v^*\| \ (\textit{according to the triangle inequality}) \\
&= \|T(v^*) - T(v_{n-1})\| + \|v_n - v^*\| \\
&\leq \beta\|\underbrace{v^* - v_{n-1}}_{\to 0}\| + \|\underbrace{v^n - v^*}_{\to 0}\|
\end{aligned}$$

Since v^* is the limit of v_n, i.e., $\lim_{n\to\infty} \|\vec{v}_n - \vec{v^*}\| = 0$ hence

$$\|T\vec{v^*} - \vec{v^*}\| = 0.$$

Thus, v^* is a fixed point of the operator T.

Uniqueness of $\vec{v^*}$
Assume that $T(v_1) = v_1$, and $T(v_2) = v_2$, and $v_1 \neq v_2$. Then

$$\|v_1 - v_2\| = \|T(v_1) - T(v_2)\| \leq \beta\|v_1 - v_2\|$$

Hence, this is in contradiction to $\beta < 1$. Therefore, v^* is unique. □

4.4.3 The Dynamic Programming Operators

We next define the basic Dynamic Programming operators, and show that they are in fact contraction operators.

For a fixed stationary policy $\pi : S \to A$, define the fixed policy DP operator $T^\pi : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$ as follows: For any $V = (V(s)) \in R^{|S|}$,

$$(T^\pi(V))(s) = r(s, \pi(s)) + \gamma \sum\nolimits_{s' \in S} p(s'|s, \pi(s))V(s'), \quad s \in S$$

In our column-vector notation, this is equivalent to $T^\pi(V) = r^\pi + \gamma P^\pi V$.

Similarly, define the discounted-return **Dynamic Programming Operator** $T^* : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$ as follows: For any $V = (V(s)) \in R^{|S|}$,

$$(T^*(V))(s) = \max_{a \in A} \left\{ r(s, a) + \gamma \sum\nolimits_{s' \in S} p(s'|s, a)V(s') \right\}, \quad s \in S$$

We note that T^π is a linear operator, while T^* is generally non-linear due to the maximum operation.

Let $||V||_\infty \stackrel{\Delta}{=} \max_{s \in S} |V(s)|$ denote the max-norm of V. Recall that $0 < \gamma < 1$.

Theorem 4.5 (**Contraction property**). *The following statements hold:*

1. *T^π is a γ-contraction operator with respect to the max-norm, namely $||T^\pi(V_1) - T^\pi(V_2)||_\infty \leq \gamma||V_1 - V_2||_\infty$ for all $V_1, V_2 \in R^{|S|}$.*

2. *Similarly, T^* is an γ-contraction operator with respect to the max-norm.*

Proof. 1. Fix V_1, V_2. For every state s,

$$\begin{aligned}|(T^\pi(V_1))(s) - (T^\pi(V_2))(s)| &= \left|\gamma \sum_{s' \in S} p(s'|s, \pi(s))[V_1(s') - V_2(s')]\right| \\ &\leq \gamma \sum_{s' \in S} p(s'|s, \pi(s)) \left|V_1(s') - V_2(s')\right| \\ &\leq \gamma \sum_{s' \in S} p(s'|s, \pi(s)) \left\|V_1 - V_2\right\|_\infty = \gamma \left\|V_1 - V_2\right\|_\infty .\end{aligned}$$

Since this holds for every $s \in S$ the required inequality follows.

2. The proof here is more intricate due to the maximum operation. As before, we need to show that $|T^*(V_1)(s) - T^*(V_2)(s)| \leq \gamma\|V_1 - V_2\|_\infty$. Fixing the state s, we consider separately the positive and negative parts of the absolute value:

(a) $T^*(V_1)(s) - T^*(V_2)(s) \leq \gamma\|V_1 - V_2\|_\infty$: Let $\bar{a}$ denote an action that attains the maximum in $T^*(V_1)(s)$, namely $\bar{a} \in \arg\max_{a \in A} \left\{r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a)V_1(s')\right\}$. Then

$$T^*(V_1)(s) = r(s, \bar{a}) + \gamma \sum\nolimits_{s' \in S} p(s'|s, \bar{a})V_1(s')$$

$$T^*(V_2)(s) \geq r(s, \bar{a}) + \gamma \sum\nolimits_{s' \in S} p(s'|s, \bar{a})V(s')$$

Since the same action $\bar{a}$ appears in both expressions, we can now continue to show the inequality (a) similarly to 1.

(b) $T^*(V_2)(s) - T^*(V_1)(s) \leq \gamma\|V_1 - V_2\|_\infty$: Follows symmetrically to (a).

The inequalities (a) and (b) together imply that $|T^*(V_1)(s) - T^*(V_2)(s)| \leq \gamma\|V_1 - V_2\|_\infty$. Since this holds for any state s, it follows that $||T^*(V_1) - T^*(V_2)||_\infty \leq \gamma\|V_1 - V_2\|_\infty$.

□

4.5 Proof of Bellman's Optimality Equation

We prove in this section Theorem 4.1, which is restated here:

Theorem (Bellman's Optimality Equation)**.** *The following statements hold:*

1. *V^* is the unique solution of the following set of (nonlinear) equations:*

$$V(s) = \max_{a \in A} \left\{r(s,a) + \gamma \sum\nolimits_{s' \in S} p(s'|s,a)V(s')\right\}, \quad s \in S. \tag{4.4}$$

2. *Any stationary policy π^* that satisfies*

$$\pi^*(s) \in \arg\max_{a \in A} \left\{ r(s,a) + \gamma \sum\nolimits_{s' \in S} p(s'|s,a) V(s') \right\},$$

is an optimal policy (for any initial state $s_0 \in S$).

We observe that the Optimality equation in part 1 is equivalent to $V = T^*(V)$ where T^* is the optimal DP operator from the previous section, which was shown to be a contraction operator with coefficient γ. The proof also uses the value iteration property of Theorem 4.2, which is proved in the next section.

Proof of Theorem 4.1: We prove each part.

1. As T^* is a contraction operator, existence and uniqueness of the solution to $V = T^*(V)$ follows from the Banach fixed point theorem. Let $\hat{V}$ denote that solution. It also follows by that theorem that $(T^*)^n(V_0) \to \hat{V}$ for any V_0. But in Theorem 4.2 we show that $(T^*)^n(V_0) \to V^*$, hence $\hat{V} = V^*$, so that V^* is indeed the unique solution of $V = T^*(V)$.

2. By definition of π^* we have

$$T^{\pi *}(V^*) = T^*(V^*) = V^*,$$

where the last equality follows from part 1. Thus the optimal value function satisfied the equation $T^{\pi *}V^* = V^*$. But we already know (from Prop. 4.2) that $V^{\pi *}$ is the unique solution of that equation, hence $V^{\pi *} = V^*$. This implies that π^* achieves the optimal value (for any initial state), and is therefore an optimal policy as stated.

□

4.6 Value Iteration

The value iteration algorithm allows to compute the optimal value function V^* iteratively to any required accuracy. The Value Iteration algorithm (Algorithm 4.2) can be stated as follows:

1. Start with any initial value function $V_0 = (V_0(s))$.

2. Compute recursively, for $n = 0, 1, 2, \ldots$,

$$V_{n+1}(s) = \max_{a \in A} \sum\nolimits_{s' \in S} p(s'|s,a)[r(s,a,s') + \gamma V_n(s')], \quad \forall s \in S.$$

3. Apply a stopping rule obtain a required accuracy (see below).

In terms of the DP operator T^*, value iteration is simply stated as:

$$V_{n+1} = T^*(V_n), \quad n \geq 0.$$

Note that the number of operations for each iteration is $O(|A| \cdot |S|^2)$. Theorem 4.2 states that $V_n \to V^*$, exponentially fast. The proof follows.

4.6.1 Error bounds and stopping rules:

While we showed an exponential convergence rate, it is important to have a criteria that would depend only on the observed quantities.

Lemma 4.3. *If* $\|V_{n+1} - V_n\|_\infty < \epsilon \cdot \frac{1-\gamma}{2\gamma}$ *then* $\|V_{n+1} - V^*\|_\infty < \frac{\varepsilon}{2}$ *and* $|V^{\pi_{n+1}} - V^*| \leq \varepsilon$, *where* π_{n+1} *is the greed policy w.r.t.* V_{n+1}.

Proof. Assume that $\|V_{n+1} - V_n\| < \varepsilon \cdot \frac{1-\lambda}{2\lambda}$, and we show that $\|V^{\pi_{n+1}} - V^*\| < \varepsilon$, which would make the policy π_{n+1} ε-optimal. We bound the difference between $V^{\pi_{n+1}}$ and V^*. (All the norms are max-norm.) We consider the following:

$$\|V^{\pi_{n+1}} - V^*\| < \|V^{\pi_{n+1}} - V_{n+1}\| + \|V_{n+1} - V^*\| \tag{4.5}$$

We now bound each part of the sum separately:

$$\begin{aligned} \|V^{\pi_{n+1}} - V_{n+1}\| &= \|T^{\pi_{n+1}}(V^{\pi_{n+1}}) - V_{n+1}\| \qquad (\textit{because } V^{\pi_{n+1}} \textit{ is the fixed point of } T^{\pi_{n+1}}) \\ &\leq \|T^{\pi_{n+1}}(V^{\pi_{n+1}}) - T^*(V_{n+1})\| + \|T^*(V_{n+1}) - V_{n+1}\| \end{aligned}$$

Since π_{n+1} is maximal over the actions using V_{n+1}, it implies that $T^{\pi_{n+1}}(V_{n+1}) = T^*(V_{n+1})$ and we conclude that:

$$\begin{aligned} \|V^{\pi_{n+1}} - V_{n+1}\| &\leq \|T^{\pi_{n+1}}(V^{\pi_{n+1}}) - T^{\pi_{n+1}}(V_{n+1})\| + \|T^*(V_{n+1}) - T^*(V_n)\| \\ &\leq \gamma\|V^{\pi_{n+1}} - V_{n+1}\| + \gamma\|V_{n+1} - V_n\| \end{aligned}$$

Rearranging, this implies that,

$$\|V^{\pi_{n+1}} - V_{n+1}\| \leq \frac{\gamma}{1-\gamma}\|V_{n+1} - V_n\| < \frac{\gamma}{1-\gamma} \cdot \epsilon \cdot \frac{1-\gamma}{2\gamma} = \frac{\epsilon}{2}$$

For the second part of the sum we derive similarly that:

$$\|V_{n+1} - V^*\| \leq \frac{\gamma}{1-\gamma}\|V_{n+1} - V_n\| < \frac{\gamma}{1-\gamma} \cdot \epsilon \cdot \frac{1-\gamma}{2\gamma} = \frac{\epsilon}{2}$$

Returning to inequality 4.5, it follows:

$$\|V^{\pi_{n+1}} - V^*\| \leq \frac{2\gamma}{1-\gamma}\|V_{n+1} - V_n\| < \epsilon$$

Therefore the selected policy π_{n+1} is ϵ-optimal. □

4.7 Policy Iteration

The policy iteration algorithm, introduced by Howard (1960), computes an optimal policy π^* in a finite number of steps. This number is typically small (on the same order as $|S|$).

The basic principle behind Policy Iteration is Policy Improvement. Let π be a stationary policy, and let V^π denote its value function. A stationary policy $\bar{\pi}$ is called π- improving if it is a greedy policy with respect to V^π, namely

$$\bar{\pi}(s) \in \arg\max_{a \in A} \left\{ r(s,a) + \gamma \sum\nolimits_{s' \in S} p(s'|s,a) V^\pi(s') \right\}, \quad s \in S.$$

Lemma 4.4 (Policy Improvement)**.** *We have $V^{\bar{\pi}} \geq V^\pi$ (component-wise), and equality holds if and only if π is an optimal policy.*

Proof. Observe first that

$$V^\pi = T^\pi(V^\pi) \leq T^*(V^\pi) = T^{\bar{\pi}}(V^\pi)$$

The first equality follows since V^π is the value function for the policy π, the inequality follows because of the maximization in the definition of T^*, and the last equality by definition of the improving policy $\bar{\pi}$.

It is easily seen that T^π is a monotone operator (for any policy π), namely $V_1 \leq V_2$ implies $T^\pi(V_1) \leq T^\pi(V_2)$. Applying $T_\gamma^{\bar{\pi}}$ repeatedly to both sides of the above inequality $V^\pi \leq T^{\bar{\pi}}(V^\pi)$ therefore gives

$$V^\pi \leq T^{\bar{\pi}}(V^\pi) \leq (T^{\bar{\pi}})^2(V^\pi) \leq \cdots \leq \lim_{n \to \infty} (T^{\bar{\pi}})^n(V^\pi) = V^{\bar{\pi}},$$

where the last equality follows by value iteration. This establishes the first claim. The equality claim is left as an **exercise**. □

The policy iteration algorithm performs successive rounds of policy improvement, where each policy π_{k+1} improves the previous one π_k. Since the number of stationary policies is bounded, so is the number of strict improvements, and the algorithm must terminate with an optimal policy after a finite number of steps.

In terms of computational complexity, Policy Iteration requires $O(|A| \cdot |S|^2 + |S|^3)$ operations per step, with the number of steps being typically small. In contrast, Value Iteration requires $O(|A| \cdot |S|^2)$ per step, but the number of required iterations may be large, especially when the discount factor γ is close to 1.

4.8 Some Variants on Value Iteration and Policy Iteration

4.8.1 Value Iteration - Gauss Seidel Iteration

In the standard value iteration: $V_{n+1} = T^*(V_n)$, the vector V_n is held fixed while all entries of V_{n+1} are updated. An alternative is to update each element $V_n(s)$ of that vector as to $V_{n+1}(s)$ as soon as the latter is computed, and continue the calculation with the new value. This procedure is guaranteed to be "as good" as the standard one, in some sense, and often speeds up convergence.

4.8.2 Asynchronous Value Iteration

Here, in each iteration $V_n \Rightarrow V_{n+1}$, only a subset of the entries of V_n (namely, a subset of all states) is updated. It can be shown that if each state is updated infinitely often, then $V_n \to V^*$. Asynchronous update can be used to focus the computational effort on "important" parts of a large-state space.

4.8.3 Modified (a.k.a. Generalized or Optimistic) Policy Iteration

This scheme combines policy improvement steps with value iteration for policy evaluation. This way the requirement for exact policy evaluation (computing $V^{\pi_k} = (I - \gamma P^{\pi_k})^{-1} r^{\pi_k}$) is avoided.

The procedure starts with some initial value vector V_0, and iterates as follows:

- Greedy policy computation:

  Compute $\pi_k \in \arg\max_\pi T^\pi(V_k)$, a greedy policy with respect to V_k.

- Partial value iteration:

  Perform m_k steps of value iteration, $V_{k+1} = (T_\gamma^{\pi_k})^{m_k}(V_k)$.

This algorithm guarantees convergence of V_k to V^*.

4.9 A Comparison between VI and PI Algorithms

In this section we will compare the convergence rate of the VI and PI algorithms. We show that, assuming that the two algorithms begin with the same approximated value, the PI algorithm converges in less iterations.

Theorem 4.6. *Let $\{VI_n\}$ be the sequence of values created by the VI algorithm (where $VI_{n+1} = T^*(VI_n)$) and let $\{PI_n\}$ be the sequence of values created by PI algorithm, i.e., $PI_n = V^{\pi_n}$. If $VI_0 = PI_0$, then for all n we have $VI_n \leq PI_n \leq V^*$.*

Proof. The proof is by induction on n.
Induction Basis: By construction $VI_0 = PI_0$. $PI_0 = V^{\pi_0}$, and therefore it is clearly bounded by V^*.
Induction Step: Assume that $VI_n \leq PI_n$. For VI_{n+1} we have,

$$VI_{n+1} = T^*(VI_n) = T^{\pi'}(VI_n),$$

where π' is the greedy policy w.r.t. VI_n, i.e., $\pi'(s) \in \arg\max_{a \in A}\{r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) VI_n(s')\}$.

Since $VI_n \leq PI_n$, and $T^{\pi'}$ is monotonic it follows that:

$$T^{\pi'}(VI_n) \quad \leq \quad T^{\pi'}(PI_n)$$

Since T^* is taking the maximum over all policies:

$$T^{\pi'}(PI_n) \quad \leq \quad T^*(PI_n)$$

The policy determined by PI algorithm in iteration $n+1$ is π_{n+1} and we have:

$$T^*(PI_n) = T^{\pi_{n+1}}(PI_n)$$

From the definition of π_{n+1}, we have

$$T^{\pi_{n+1}}(PI_n) \quad \leq \quad V^{\pi_{n+1}} = PI_{n+1}$$

Therefore, $VI_{n+1} \leq PI_{n+1}$. Since $PI_{n+1} = V^{\pi_{n+1}}$, it implies that $PI_{n+1} \leq V^*$. □