3 Dynamic Programming – Infinite Horizon

3.1 Performance Criteria

We next consider the case of infinite time horizon, namely $\mathbf{T} = \{0,1,2,\ldots,\}$. The importance of the infinite horizon model relies on the following observations:

1. In many problems, a specific finite time horizon is not easily specified, and the infinite horizon formulation is more natural.
2. More importantly – stationary problems with infinite time horizon lead to optimal stationary strategies, which offer great simplicity.

There are several possible performance criteria for infinite-horizon problems. The more common ones are:

1. *Total cost:*

$$J^{\pi,s} = E^{\pi,s}\Big(\sum_{t=0}^{\infty} R_t\Big) = E^{\pi,s}\Big(\sum_{t=0}^{\infty} r_t(s_t, a_t, s_{t+1})\Big)$$

2. *Discounted cost*:

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

   where $0 < \gamma < 1$ is the *discount factor*.

3. *Average cost:*

$$J^{\pi,s} = \liminf_{N\to\infty} E^{\pi,s}\Big(\tfrac{1}{N}\sum_{t=0}^{N-1} R_t\Big)$$

* Note: we use the common terminology of "cost" even though here we consider rewards that we wish to maximize. The terms "total return" or "total reward" etc. can be used instead, but are less common.

It is obvious that the total cost might diverge (go to infinity, or even not converge at all if both negative and positive rewards are possible). Therefore, further model assumptions are required to ensure that the optimization problem is well defined in that case.

The discounted cost is the "best behaved" – discounting ensures convergence (at least when the one-stage rewards are bounded), and therefore no additional assumptions are required.

The average cost requires more sophisticated analysis, related to long-term properties of Markov chains. It will not be treated in this chapter.

3.2 Discounted Cost

We consider first the discounted cost criterion:

$$J_\gamma^\pi(s) = E^\pi\Big(\sum_{t=0}^{\infty} \gamma^t R_t \mid s_0 = s\Big)$$

where $0 < \gamma < 1$. The sum converges when R_t is bounded. This is trivially ensured in the case of finite state and action spaces that we consider here.

We further consider the *stationary* problem:

$$p_t(s' \mid s,a) \equiv p(s' \mid s,a) \text{ and } r_t(s,a,s') \equiv r(s,a,s') \text{ for all } t \geq 0 .$$

3.2.1 The Basic Operators

Denote

$$V^\pi(s) = J_\gamma^\pi(s)$$

$$V^*(s) = \sup_\pi V^\pi(s)$$

where the supremum is taken over all (general) policies.

Let π denote a stationary policy, namely $\pi : S \to A$ and $a_t = \pi(s_t)$. Let $\mathbb{R}^S$ denote the space of functions $V : S \to \mathbb{R}$. Note that V can also be viewed as an $|S|$-dimensional vector.

Define the following operators over $\mathbb{R}^S$:

$$T_\gamma^\pi : \quad (T_\gamma^\pi V)(s) = \sum\nolimits_{s' \in S} p(s' \mid s, \pi(s))[r(s,\pi(s),s') + \gamma V(s')]$$

$$T_\gamma^* : \quad (T_\gamma^* V)(s) = \max_{a \in A} \sum\nolimits_{s' \in S} p(s' \mid s,a)[r(s,a,s') + \gamma V(s')]$$

Let $\| V \|_\infty \triangleq \max_{s \in S} | V(s) |$ denote the max-norm of V.

Theorem 1 (Contraction property)

(i) T_γ^π is a γ-contraction operator with respect to the max-norm, namely

$$\| T_\gamma^\pi V_1 - T_\gamma^\pi V_2 \|_\infty \le \gamma \| V_1 - V_2 \|_\infty \text{ for all } V_1, V_2 \in \mathbb{R}^S$$

(ii) Similarly, T_γ^* is an γ-contraction operator with respect to the max-norm.

Proof: (i) For every s,

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

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

(ii) Home exercise (a bit harder!). □

We next quote the following basic property of contraction operators.

Theorem (The Banach fixed point theorem). Let V be a Banach space (namely, a complete normed linear space) , and let $T : \mathrm{V} \to \mathrm{V}$ be a contraction operator, namely there exists $\beta \in (0,1)$ so that $\| TV_1 - TV_2 \| \le \beta \| V_1 - V_2 \|$ for all $V_1, V_2 \in \mathrm{V}$. Then

(C1) The equation $TV = V$ has a unique solution $V^* \in \mathrm{V}$.

(C2) For any $V_0 \in \mathrm{V}$, $\lim_{n\to\infty} T^n V_0 = V^*$.

Proof:

(C1) *Uniqueness:* Let V_1 and V_2 be two solutions of $TV = V$, then

$$\| V_1 - V_2 \| = \| TV_1 - TV_2 \| \le \beta \| V_1 - V_2 \|,$$

which implies that $\| V_1 - V_2 \| = 0$, or $V_1 = V_2$.

Existence (outline): (i) show that $V_n \triangleq T^n V_0$ (with V_0 arbitrary) is a Cauchy sequence. (ii) Since the space V is complete by assumption, this implies that V_n converges to some $V^* \in \mathrm{V}$. (iii) Now show that V^* satisfies the equation $TV = V$.

(C2) We have just shown that, for any V_0, $V_n \triangleq T^n V_0$ converges to a solution of $TV = V$, and that solution was shown before to be unique. □

3.2.2 Value Iteration with a Fixed Policy

Proposition 2 Let π be a *stationary* policy, and let V^{π} denote its value function. Then V^{π} is the unique solution of $V \equiv T^{\pi}V$. More explicitly, we have

$$V(s)=\sum\nolimits_{s'\in S} p(s'\,|\,s,\pi(s))[r(s,\pi(s),s')+\gamma V(s')]\,, \quad s\in S \quad (*)$$

Proof: That V^{π} satisfies the equation $V \equiv T_{\gamma}^{\pi}V$ follows similarly to the finite horizon case. Uniqueness of the solution follows since T_{γ}^{π} is a contraction, see property (C1) of the fixed-point theorem. □

It is important to note that (*) is just a set of $|S|$ *linear* equations. If we define the vector r^{π} via $r^{\pi}(s)=\sum\nolimits_{s'} p(s'\,|\,s,\pi(s))\,r(s,\pi(s),s')$ and the transition matrix P^{π} via $P^{\pi}(s'\,|\,s)=p(s'\,|\,s,\pi(s))$, then we can write this set of equations as $V=r^{\pi}+\gamma P^{\pi}V$, with solution $V^{\pi}=(I-\gamma P^{\pi})^{-1}r^{\pi}$.

Proposition 3 For any initial value V_0, define recursively $V_n=T_{\gamma}^{\pi}V_{n-1}$. Then $V_n \to V^{\pi}$.

Proof: Since $V_n=(T_{\gamma}^{\pi})^{n}V_0$, the required convergence follows from property (C2) of the contraction T_{γ}^{π} together with the previous proposition. □

In fact, it is easy to verify that

$$\begin{aligned} V_n(s) &= E^{\pi}\left(R_0+\gamma V_{n-1}(s_1)\,|\,s_0=s\right) \\ &= \ldots = E^{\pi}\left(\sum\nolimits_{t=0}^{n-1}\gamma^{t}R_t+\gamma^{n}V_0(s_N)\,|\,s_0=s\right) \end{aligned}$$

That is, V_n is the n-step discounted cost with terminal cost function V_0 (under the policy π). Since $\gamma<1$, it is easy to see directly that $V_n \to V^{\pi}$.

In summary:

- Proposition 2 allows to compute V^{π} by solving a set of $|S|$ linear equations.
- Proposition 3 computes V^{π} by an infinite recursion.

3.2.3 Bellman's Optimality Equation

Recall that $V^*(s) = \sup_{\pi\in\Pi} V^\pi(s)$ is the optimal value function (with respect to the γ-discounted cost criterion).

Theorem 4 (Optimality Equation)
(i) V^* is the unique solution of $V = T_\gamma^* V$, namely of the following set of (nonlinear) equations:

$$V(s) = \max_{a\in A} \sum\nolimits_{s'\in S} p(s'\,|\,s,a)[r(s,a,s') + \gamma V(s')] , \qquad s \in S .$$

(ii) Any stationary policy π^* that satisfies

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

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

For the proof, we first establish the following result.

Theorem 5 (Value Iteration)
Starting with an arbitrary $V_0 \in \mathbb{R}^S$, define recursively $V_{n+1} = T_\gamma^* V_n$, namely

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

Then $\lim_{n\to\infty} V_n = V^*$, where the rate of convergence is exponential.
A short version of this claim: $(T_\gamma^*)^n V_0 \to V_\alpha^*$, for any V_0.

Proof: Using our previous results on the finite-horizon problem, it follows immediately that

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

This may be verified to be within a margin of $\gamma^n (R_{\max}/(1-\gamma) + \| V_0 \|_\infty)$ from the optimal value $V_\gamma^*(s)$. As $\gamma < 1$, this implies that V_n converges to V_γ^* at an exponential rate. □

Proof of Theorem 4:

(i) Recall that T_γ^* is a contraction operator. This implies the existence and uniqueness of the solution to $V = T_\gamma^* V$. Let $\hat{V}$ denote that solution. The contraction property also implies that $(T_\gamma^*)^n V_0 \to \hat{V}$ for any V_0. But in Theorem 5 we showed that $(T_\gamma^*)^n V_0 \to V^*$, hence $\hat{V} = V^*$ and V^* is the unique solution of $V = T_\gamma^* V$.

(ii) By definition of π^* we have

$$T_\gamma^{\pi^*} V^* = T_\gamma^* V^* = V^*$$

where the last equality follows from (i). Thus the optimal value function satisfied the equation $T_\gamma^{\pi^*} V^* = V^*$. But we already know (from Prop. 2) that V^{π^*} 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. □

The optimality equation provides a useful characterization of the optimal value function and of the optimal control policy. However, this equation is non-linear in general (due to the required maximization), and cannot be solved directly. Several methods have been devised for the solution of this equation, with the most basic ones being:

- *Value Iteration*
- *Policy Iteration*
- *Linear Programming*

We proceed to explain these methods.

3.2.4 Value Iteration

The *Value Iteration algorithm* is already defined in Theorem 5, and repeated here:

- Start with any initial value function V_0.
- Compute recursively $V_{n+1} = T_\gamma^* V_n$, namely

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

As $(T_\gamma^*)^n V_0 \to V^*$, we can compute V^* to any accuracy. Note that the number of operations for each iteration here is $O(|A| \cdot |S|^2)$.

<u>Error bounds and stopping rules:</u>

It is important to have an on-line criterion for the accuracy of the n-the step solution V_n. We quote some basic bounds without proof.

(a) The distance of V_n from the optimal solution is upper bounded as follows:

$$\| V_n - V^* \|_\infty \leq \tfrac{\gamma}{1-\gamma} \| V_n - V_{n-1} \|_\infty$$

Note that the right-hand side also decays exponentially (with rate γ). This enables to compute the value function to within any required accuracy: To ensure $\| V_n - V^* \|_\infty \leq \varepsilon$, we need to verify that $\| V_n - V_{n-1} \|_\infty \leq \frac{1-\gamma}{\gamma} \varepsilon$.

(b) If $\| V - V^* \|_\infty \leq \varepsilon$, then any stationary policy π that is *greedy* with respect to V, namely satisfies

$$\pi(s) \in \arg\max_{a \in A} \sum\nolimits_{s' \in S} p(s' | s, a)[r(s, a, s') + \gamma V_n(s')]$$

is 2ε-optimal, namely $\| V^\pi - V^* \|_\infty \leq 2\varepsilon$.
This enable to obtain a 2ε-optimal policy from an ε-approximation of V^*.

More refined error bounds can be found in the texts on Dynamic Programming.

3.2.5 Policy Iteration

This procedure (by Howard, 1960) computes V^* and 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 be its value function. A stationary policy $\bar{\pi}$ is called π-*improving* if it is a greedy policy with respect to V^π, namely

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

Lemma 6 (Policy Improvement)
$V^{\bar{\pi}} \geq V^\pi$ (component-wise), and equality holds *if and only if* π is optimal.

Proof: Observe first that

$$V^\pi = T_\gamma^\pi V^\pi \leq T_\gamma^* V^\pi = T_\gamma^{\bar{\pi}} V^\pi$$

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

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

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

The last lemma leads to a finite algorithm for computing an optimal policy π^*.

The Policy Iteration Algorithm:

```
0. Initialization: choose some stationary policy $\pi_0$.
For $k = 0,1,\ldots$:
1. Policy evaluation: compute $V^{\pi_k}$,
    e.g. by using the explicit formula $V^{\pi_k} = (I - \gamma P^{\pi_k})^{-1} r^{\pi_k}$.
2. Policy Improvement: Compute $\pi_{k+1}$, a greedy policy with respect to $V^{\pi_k}$.
3. Stop if $V^{\pi_{k+1}} = V^{\pi_k}$ (or if $T_\gamma^* V^{\pi_k} = V^{\pi_k}$), else repeat.
```

If follows from Lemma 6 that each policy π_{k+1} is strictly better than the previous one (unless π_k is already optimal). Since the number of stationary deterministic policies is finite, this algorithm must end with an optimal policy in 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.

3.2.6 Some Variants on Value & Policy Iteration

A. Value Iteration – Gauss Seidel Iteration

In the standard value iteration: $V_{n+1} = T_\gamma^* V_n$, 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.

B. Asynchronous Value Iteration

Here, in each iteration $V_n \mapsto V_{n+1}$, 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.

C. Modified (/Generalized/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 $\pi_k \in \arg\max_\pi T_\gamma^\pi V_k$, 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^*.

Note that extreme values of m_k (which?) reduce this algorithm to the standard Value Iteration or Policy Iteration.

3.2.7 Linear Programming

An alternative approach to value and policy iteration is the linear programming method. Here the optimal control problem is formulated as a linear program (LP), which can be solved efficiently using standard LP solvers. There are two formulations: *primal* and *dual*. As this method is less related to learning we will only sketch it briefly.

a. The Primal LP

Recall that V^* satisfies the optimality equations:

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

Claim: V^* is the smallest function (component-wise) that satisfies the following set of inequalities:

$$v(s) \geq r(s,a) + \gamma \sum\nolimits_{s' \in S} p(s' | s,a) v(s'), \quad \forall s,a \qquad (*)$$

Proof: Suppose $v = (v(s))$ satisfies $(*)$. That is, $v \geq T_\gamma^\pi v$ for every stationary policy π. Then by the monotonicity of T_γ^π,

$$v \geq T_\gamma^\pi v \;\Rightarrow\; T_\gamma^\pi v \geq (T_\gamma^\pi)^2 v \;\Rightarrow\; \ldots \Rightarrow\; (T_\gamma^\pi)^k v \geq (T_\gamma^\pi)^{k+1} v$$

so that

$$v \geq T_\gamma^\pi v \geq (T_\gamma^\pi)^2 v \geq \ldots \geq \lim_{n \to \infty} (T_\gamma^\pi)^n v = V^\pi$$

Now, if we take π as the optimal policy we obtain $v \geq V^*$ (component-wise). □

It follows from the last claim that V^* is the solution of the following linear program:

Primal LP: $\displaystyle \min_{(v(s))} \sum_s v(s)$

subject to $(*)$

Note that the number of inequality constraints is $N_S N_A$.

b. The Dual LP

Given an LP

$$\min_{x \geq 0} b^T x, \quad \text{s.t. } Ax \geq c$$

its dual LP is defined as:

$$\max_{y \geq 0} c^T y, \quad \text{s.t. } A^T y \leq b$$

We note that the dual LP is obtained by noting that the optimal value of the primal equals:

$$\min_{x\geq 0}\max_{y\geq 0}\left\{b^T x + y^T(c - Ax)\right\} = \max_{y\geq 0}\min_{x\geq 0}\left\{c^T y + x^T(b - Ay)\right\}$$

The two LPs have the same optimal value, and the solution of one can be obtained from that of the other.

The dual of our Primal LP turns out to be:

Primal LP: $\max_{(f_{s,a})} \sum_{s} f_{s,a} r(s,a)$

subject to:

$$f_{s,a} \geq 0 \quad \forall s,a$$

$$\sum_{s,a} f_{s,a} = \tfrac{1}{1-\gamma}$$

$$p_0(s') + \gamma \sum_{s,a} p(s'\,|\,s,a) f_{s,a} = \sum_{s} f_{s,a} \quad \forall s' \in S$$

where $p_0 = (p_0(s'))$ is any probability vector (usually take as a 0/1 vector).

Interpretation:

- The variables $f_{s,a}$ correspond to the "state action frequencies" (for a given policy):

$$f_{s,a} \sim E(\sum_{t=0}^{\infty} \gamma^t I_{\{s_t=s,a_t=a\}}) = \sum_{t=0}^{\infty} \gamma^t P(s_t = s, a_t = a)$$

  and $p_0(s') \sim p(s_0 = s')$ is the initial state distribution.

- It is easily to see that the discounted cost can be written in terms of $f_{s,a}$ as:

$$\sum_{s} f_{s,a} r(s,a)$$

  which is to be maximized.

- The above constraints easily follow from the definition of $f_{s,a}$.

Further comments:

- The optimal policy can by obtained directly from the solution of the dual using:

$$\pi(a\,|\,s) = \frac{f_{s,a}}{f_s} \equiv \frac{f_{s,a}}{\sum_a f_{s,a}}$$

  This policy can be stochastic if the solution to the LP is not unique. However, it will be deterministic even in that case if we choose f as an *extreme* solution of the LP.

- The number of constraints in the dual is $N_S N_A + (N_S + 1)$. However, the inequality constraints are simpler than in the primal.

3.3 Stochastic Shortest-Path (SSP) Problems

This class of problems generalizes the standard shortest path problem, in which the goal is to find the shortest path to a given goal state. SSP considers the (undiscounted) total cost, with the number of steps not bounded a-priori. However, the process effectively halts when it reaches a given "goal" state.

Consider then the stationary MDP model, with (undiscounted) total cost function:

$$V^{\pi}(s) = E^{\pi}\Big(\sum_{t=0}^{\infty} R_t \mid s_0 = s\Big)$$

We assume that there exists a specific state, the *goal* or *terminal state*, denoted '0'. This state has no actions, and is

(a) Absorbing: $p(0 \mid 0) = 1$.

(b) Costless: $r(0) = 0$.

To ensure that the total cost is well defined and meaningful, we will add some assumptions on the model that ensure that the terminal state can and will be reached. One possible set of assumptions is the following:

Assumption A

(A1) For every state s there exists a stationary policy π so that state 0 is reached from s with positive probability.

(A2) $V^{\pi}(s) = -\infty$ for any s and π which do *not* satisfy the above.

Remarks:

- Essentially – (A1) ensures we *can* reach state 0, and (A2) ensures we *want* to reach it.
- (A1) can be seen to be equivalent to existence of a *fixed* policy π under which state 0 is reached from every state s with positive probability. Such a policy is called *proper*.
- In the deterministic-transitions case, (A1) means there is a path to state 0 from every state s. (A2) then means that every non-terminating path has infinite cost. This is equivalent to requiring that every loop (closed path) has strictly negative reward.
- (A2) can be replaced with $r(s,a) < 0$, the proofs are actually easier then.

Results: The SSP problem is harder than the discounted one, because of lack of immediate contraction properties. However, the main results do carry over – including existence of optimal stationary policies, optimality equations, value iteration, and policy iteration. The Dynamic Programming operators are the same as before, with $\gamma = 1$.

We note however that the solutions to the DP equations are unique only up to a constant vector, that is, if $(v(s))_{s\in S}$ is a solution, then so is $(v(s)+c)_{s\in S}$. Uniqueness is obtained by noting that $V('0') = 0$ holds for the absorbing state.

Relation to other problems: Both the finite-horizon problem and the infinite-horizon discounted-cost problem can be embedded within an SSP problem (exercise).

We finally note that in the deterministic version of SSP, value iteration terminates in a finite number of steps, and both backward and forward recursions are possible.