Chapter 6

Infinite Horizon Discounted Cost Problems

References: [Ber07, Vol. I, ch. 7, Vol II ch. 1]

6.1 Introduction

[1]In this chapter, we start our investigation of infinite horizon problems. There are several reasons for considering optimal control problems with an infinite horizon.

- The number of decision stages is really infinite, or at least a large number which is not precisely known and distant into the future, in which case the infinite-horizon approximation is a good model.
- In contrast to our assumption in the previous chapters, the horizon length can be random, or itself subject to control (optimal stopping) with no a priori specified upper bound.
- We often approximate a large number of periods, even if the horizon is known and finite, by assuming an infinite number of periods, and hope that this assumption will simplify the solution. Indeed, even if the general theory becomes more involved, the solution obtained often is simpler and has important computational and conceptual advantages: in particular, the optimal policy is often stationary (the same function at every stage). For example, recall our discussion at the end of chapter 4 on linear quadratic Gaussian problems, where we mentioned that the optimal infinite horizon controller gain is constant and requires solving only one algebraic Riccati equation, whereas for the optimum finite horizon problem one must compute and/or store a different gain matrix for each time step. If the number of stages is large, we saw by an example that the difference in performance is typically negligible.

[1] This version: October 11 2009.

The main issue arising when one considers infinite horizon problems is that the backward DP recursion needs to be replaced by something else: indeed, it would require initialization at $k = \infty$ and backward recursion for an infinite number of steps to compute the value function J_0^* ! The theoretical tool that replaces the DP algorithm is a steady-state version of the DP recursion, called the Bellman equation (for continuous-time system, the corresponding equation is called the Hamilton-Jacobi-Bellman equation, a terminology which is sometimes also used for discrete time systems, and even in some cases for the DP recursion in finite horizon problems). There are unfortunately a variety of mathematical technicalities that force us to verify, depending on our problem assumptions, that Bellman's equation indeed holds.

For computations, the direct generalization of the DP algorithm to the infinite horizon problem is called *value iteration*. There are other ways of solving the Bellman's equation as well, and we introduce another well-known method in chapter 7, called *policy iteration*. In this chapter, we first give a brief overview of some of the main classes of infinite horizon stochastic optimal control problems, which usually require different sets of conditions and different proof techniques to insure the validity of Bellman's equation. Then we consider a class of problems with *perfect information*, a *discounted and bounded stage cost* and a *discrete state space*. This formulation has the simplest theory, still has a wide range of applications, and is often the first approach tried, perhaps by approximation, when dealing with a new problem.

6.2 Overview of Infinite Horizon Problems

When considering infinite-horizon problems, we will assume that the system is *time-homogeneous* (also called *stationary*, but this does not mean that $\{x_k\}_{k\geq 0}$ is a stationary stochastic process), i.e., the system equation, cost per stage, random disturbance statistics, state and control spaces do not change in time:

$$x_{k+1} = f(x_k, u_k, w_k),$$

where $x_k \in \mathsf{X}$, $u_k \in \mathsf{U}(x_k)$, $w_k \in \mathsf{W}$, and the distribution $P_{w_k}(\cdot|x_k, u_k)$ of the disturbances is now independent of k. Infinite-horizon problems are divided into several classes, for which the analysis techniques are often different. First consider problems where we simply try to minimize the total cost over an infinite number of stages. The cost associated to a given admissible policy $\pi = \{\mu_0, \mu_1, \ldots\}$ (i.e., $\mu_k(x) \in \mathsf{U}_k(x)$ for all k) and initial state x_0 is now

$$J_\alpha^\pi(x_0) := J_\alpha(x_0; \pi) = \lim_{N\to\infty} E^\pi \left[\sum_{k=0}^{N-1} \alpha^k c(x_k, \mu_k(x_k), w_k) \,\middle|\, x_0 \right], \tag{6.1}$$

where $0 \leq \alpha \leq 1$ is a *discount factor* (we replace lim by lim sup or lim inf if it is not known that the limit exists). Note that the cost-function is also assumed to be the same at each stage. The optimal cost J^* is defined by

$$J_\alpha^*(x) = \min_{\pi\in\Pi} J_\alpha^\pi(x), \tag{6.2}$$

where Π is the set of admissible policies. Issues immediately arise concerning the meaning of this problem, because it might well be that no matter what policy is used, the total cost (6.1) is infinite (i.e., the limit is equal to $+\infty$ or $-\infty$. For example, a one-state MDP with a single action and a nonzero reward), or even non-convergent (the limit does not exist). So different assumptions are made on the problem structure and discount factor, which aim at guaranteeing a well-defined finite cost, at least for some of the policies. Here are some of the main classes of problems studied.

1. *Discounted cost problems with bounded cost per stage.* Here $0 \leq \alpha < 1$. These problems are often found in economics applications, where $\alpha = 1/(1+\gamma)$ with $\gamma > 0$ a rate of interest or inflation rate, in operations research, robotics and artificial intelligence, where $(1-\alpha)$ might be interpreted as the probability at each stage that the system under study breaks down permanently (see e.g. [Put94, p.126] for mathematical justification of this interpretation), or purely for mathematical convenience (of course, this is usually not stated that way). The role of the discount rate is to emphasize short-term rewards vs. rewards that might be obtained in a more distant future. In the limiting case $\alpha = 0$, we are only concerned about the expected cost of the first stage $E_{w_0}[c(x_0, \mu(x_0), w_0)]$, which is a (stochastic) optimization problem. A discount factor $\alpha < 1$ is not sufficient to guarantee convergence in general. However, the limit (6.1) exists and is finite when the cost at every stage is uniformly bounded

   $$\sup_{x \in \mathsf{X}} \sup_{u \in \mathsf{U}(x)} \sup_{w} |c(x,u,w)| \leq M < \infty. \tag{6.3}$$

   Here the $\sup_w$ is the supremum over the values of w which can arise with positive probability given x, u. In this case $|J_\alpha^\pi(x_0)| < M/(1-\alpha)$ for all $x_0 \in \mathsf{X}$ and all policies π, and the cost occurring after the k^{th} step is bounded by $\alpha^k M/(1-\alpha)$. Moreover by the dominated convergence theorem we can interchange the limit and expectation in (6.1)

   $$J_\alpha^\pi(x_0) := J_\alpha(\pi, x_0) = E_{x_0}^\pi \left[\sum_{k=0}^{\infty} \alpha^k c(x_k, \mu_k(x_k), w_k) \right].$$

   We will look at this problem in more details in this chapter. Note that (6.3) is automatically satisfied if the state, control and disturbance spaces are finite (why?).

2. *Discounted and undiscounted cost problems with unbounded cost per stage.* Assumption (6.3) is quite problematic because it does not even allow us to consider the LQR problem with discount $\alpha < 1$. But if we relax this assumption the cost can become infinite under certain policies, which we need to rule out during analysis. Common additional assumptions made to analyze these problems are: assuming that all stage costs are nonnegative (called *negative* models and negative dynamic programming!) or

assuming that in all states there is an action with negative cost, or the stronger assumption that all costs are nonpositive (called *positive* models and positive dynamic programming!). The explanation for the inverted terminology is that it comes from optimistic people who are maximizing rewards instead of minimizing costs. In negative and positive dynamic programming one usually assumes $\alpha = 1$. The existence of the limit (6.1) is then guaranteed by the monotone convergence theorem.

3. *Stochastic Shortest Path Problems:* here $\alpha = 1$, but there is a special absorbing cost-free termination state. Note that this is somewhat related to the random termination interpretation above of the discounted cost problem.

4. *Average Cost Problems.* Here (6.1) is replaced by

$$\lim_{N\to\infty} \frac{1}{N} E^{\pi}\left[\sum_{k=0}^{N-1} c(x_k, \mu_k(x_k), w_k) \,\Big|\, x_0\right]. \tag{6.4}$$

   With this criterion, the controller aims at optimizing the steady-state behavior of the system. The theory for this problem is strongly related to the asymptotic theory of Markov chains. In particular, notions of stability come into play. It is the criterion optimized in the standard infinite-horizon LQG problem, and it is often used in the literature on queueing networks. Often the form of the optimal policy can be simpler than for other criteria, such as discounted cost problems. The main drawback of this criterion is that it does not take into account any transient regime, which can be important or even the most important aspect of a control problem. For example, modifying an optimal policy at a finite number of steps still yields an optimal policy in general. So this criterion has limitations in distinguishing policies, even if these have very different appeal to a decision maker.

5. *Other optimality criteria.* Other classifications and refinements are possible. The point is to be able to isolate a problem structure for which the infinite-horizon cost makes sense and which at the same time is broad enough to capture interesting applications. Puterman discusses for example additional ways of refining the average-cost optimality criterion, such as the overtaking optimality criterion and sensitive discount optimality criterion [Put94, chapter 5].

6.3 The Dynamic Programming Operator

The time-homogeneity assumption allows us to solve recursively the sequence of finite horizon problems corresponding to (6.1) as the horizon length N increases. Given a fixed N and an arbitrary bounded terminal cost, denoted $J(x)$

(instead of $c_N(x)$ previously), consider the finite horizon problem

$$J_0^*(x_0;N) = \min_{\{\mu_0,\dots,\mu_{N-1}\}} E\left[\alpha^N J(x_N) + \sum_{k=0}^{N-1} \alpha^k c(x_k,\mu_k(x_k),w_k) \,\Big|\, x_0\right].$$

The notation $J_0(x;N)$ is used to remember the fact that the problem is now parameterized by N. We wish to compute $J_0(x_0;N)$ for N increasing toward $+\infty$ in order to compute (6.1). For the N-stage finite horizon problem, the DP algorithm is

$$\begin{aligned} J_N^*(x;N) &= \alpha^N J(x) \\ J_{N-k}^*(x;N) &= \min_{u\in\mathsf{U}(x)} E\Big[\alpha^{N-k}c(x,u,w) + J_{N-k+1}(f(x,u,w)) \,\Big|\, x_{N-k} = x\Big], \end{aligned}$$

where we chose to write the DP recursion for the problem with k *remaining* stages. Another expression for J_{N-k}^* is

$$\begin{aligned} J_{N-k}^*(x;N) = \alpha^{N-k} E\Bigg[&\alpha^k J(x_N) \\ &+ \sum_{i=0}^{k-1} \alpha^i c(x_{N-k+i},\mu_{N-k+i}^*(x_{N-k+i}),w_{N-k+i}) \Big| x_{N-k} = x\Bigg]. \end{aligned}$$

Let $V_k(x) = J_{N-k}^*(x;N)/\alpha^{N-k}$, *which represents the optimal cost for a k-stage problem*, as is apparent from the previous equation. In particular, $V_k(x)$ does not depend on N. Note that $V_0(x) = J(x)$, $V_N(x) = J_0^*(x;N)$ (the quantity of interest), and from the DP algorithm we see that V_k satisfies the recursion

$$V_{k+1}(x) = \min_{u\in\mathsf{U}(x)} E\Big[c(x,u,w) + \alpha V_k(f(x,u,w)) \,\Big|\, x\Big]. \tag{6.5}$$

Hence $V_N(x)$ is the optimal cost for a finite horizon problem with horizon N. Once the recursion is initialized with $V_0 = J$, we can compute the optimal cost for any horizon using a single recursion. Hence if we have computed $J_0^*(\cdot;N) = V_N$ for a horizon of length N, we obtain $J_0^*(\cdot;N+1) = V_{N+1}$ for a horizon of length $N+1$ using (6.5), without having to restart the DP algorithm from the initial stage. It is natural to conjecture that as $k \to \infty$, V_k converges to the optimum cost (6.1).

The following notation will be used extensively in the following. The recursion step (6.5) maps a function V_k to a new function V_{k+1}. Let us redefine it as an abstract operator, denoted T, on the space of functions from X to $\mathbb{R}$

$$(TJ)(x) = \min_{u\in\mathsf{U}(x)} E\Big[c(x,u,w) + \alpha J(f(x,u,w)) \,\Big|\, x\Big]. \tag{6.6}$$

So T maps a real-valued function J defined on X to a new function TJ on X. If the system is specified in the form of a controlled Markov chain instead of the state-space form, with a countable state space and transition probabilities $p_{xy}(u)$, then (6.6) is written

$$(TJ)(x) = \min_{u \in \mathsf{U}(x)} \Big\{ \sum_{y \in \mathsf{X}} p_{xy}(u) \Big[c(x,u,y) + \alpha J(y) \Big] \Big\}.$$

On a general state space, we would write

$$(TJ)(x) = \min_{u \in \mathsf{U}(x)} \Big\{ \int_{y \in \mathsf{X}} \Big[c(x,u,y) + \alpha J(y) \Big] dP(y|x,u) \Big\}.$$

Remark that TJ is *the optimal cost function for a one-stage α-discounted problem with stage cost c and terminal cost αJ.* Similarly, for a control function $\mu : \mathsf{X} \to \mathsf{U}$, we define the operator T_μ by

$$(T_\mu J)(x) = E\Big[c(x, \mu(x), w) + \alpha J(f(x, \mu(x), w)) \Big| x \Big].$$

Hence $T_\mu J$ can be viewed as the cost associated with the control μ in a one-stage α-discounted problem with stage cost c and terminal cost αJ. We denote $T^0 J \equiv J$, and define recursively the iterates $T^k J \equiv T(T^{k-1}J), k \geq 1$. We have a similar notation for T_μ. By immediate backward induction, we have that $V_k \equiv T^k J$, so $T^k J(x)$ is the optimal cost for the k-stage, α-discounted problem with initial state x, cost per stage c and terminal cost $\alpha^k J$. Finally, $T_\mu^k J$ and $T_{\mu_0} T_{\mu_1} \ldots T_{\mu_{k-1}} J$ are the costs of the policy $\{\mu, \mu, \ldots\}$ and $\{\mu_0, \mu_1, \ldots \mu_{k-1}\}$ respectively for the same problem. We can rewrite the DP algorithm in a compact form using these operator. Equation (6.5) is

$$V_{k+1} = TV_k.$$

Moreover, if we consider a problem with finite horizon N and optimal policy $\pi = \{\mu_0^*, \ldots, \mu_{N=1}^*\}$, then μ_i^* satisfies

$$T_{\mu^*_{N-k-1}} V_k = TV_k,$$

which simply means that μ^*_{N-k-1} achieves the minimum in the DP recursion (6.5). We now prove some useful properties of the operator T. For two functions $f, g : \mathsf{X} \to \mathbb{R}$, we use the notation $f \leq g$ iff $f(x) \leq g(x)$ for all $x \in \mathsf{X}$.

Lemma 6.3.1 (monotonicity lemma)**.** *For $J, J' : \mathsf{X} \to \mathbb{R}$, if $J \leq J'$ then $TJ \leq TJ'$ and $T_\mu J \leq T_\mu J'$. Hence for all $k \geq 0$, $T^k J \leq T^k J'$ and $T_\mu^k J \leq T_\mu^k J'$.*

Proof. Immediate by definition of TJ and TJ' as cost functions for the same one-stage cost problem except for the final costs $J \leq J'$. □

Next consider the constant unit function $e : \mathsf{X} \to \mathbb{R}$, with $e(x) = 1$ for all $x \in \mathsf{X}$. The following properties follows by straightforward induction.

Lemma 6.3.2 (offset property lemma)**.** *For any $r \in \mathbb{R}$, $k \geq 0$, we have*

$$T^k(J + re) = T^k J + \alpha^k re$$
$$T_\mu^k(J + re) = T_\mu^k J + \alpha^k re.$$

These two properties hold for any value of α. Next, we turn to the crucial property which is at the origin of the simpler analysis of the convergence properties in the discounted cost case. Let $B(\mathsf{X}, \mathbb{R})$ be the set of bounded real-valued functions on $\mathbb{R}$. It turns out that T is an α-contraction on $B(\mathsf{X}, \mathbb{R})$ (note that T is not a linear or affine mapping however, although T_μ is affine).

Exercise 13. Show that if $J \in B(\mathsf{X}, \mathbb{R})$ and the bounded cost per stage assumption (6.3) is satisfied, then $TJ \in B(\mathsf{X}, \mathbb{R})$.

Lemma 6.3.3 (max-norm contraction lemma)**.** *Under the bounded cost per stage assumption (6.3), T and T_μ are α-contractions on $B(\mathsf{X}, \mathbb{R})$ for the sup-norm $\| \cdot \|_\infty$. For all $J, J' : \mathsf{X} \to \mathbb{R}$ bounded, and for all $k \geq 0$, we have*

$$\|T^k J - T^k J'\|_\infty \leq \alpha^k \|J - J'\|_\infty,$$
$$\|T^k_\mu J - T^k_\mu J'\|_\infty \leq \alpha^k \|J - J'\|_\infty.$$

Proof. Let $r = \|J - J'\|_\infty$. Then

$$J' - re \leq J \leq J' + re.$$

Then use the monotonicity lemma and the offset property to get

$$T^k J' - \alpha^k re \leq T^k J \leq TJ' + \alpha^k re.$$

In other words, $\|T^k J - T^k J'\|_\infty \leq \alpha^k r$. □

Finally, we show that starting with a bounded function J, the iterates $V_0 \equiv J, V_1 = TV_0 = TJ, \ldots, V_k = TV_{k-1} = T^k J, \ldots$, converge to the true cost, as expected. Note that the initial function J does not impact the result, as long as it is bounded. This is expected since its effect is essentially forgotten through the discounting (recall that the terminal cost in the definition of V_k as the cost of a k-stage problem is $\alpha^k J$). Also, performing these iterations gives us an algorithm to compute the optimal cost function, called the *value iteration* algorithm.

Lemma 6.3.4 (convergence of the DP algorithm)**.** *For any $J : \mathsf{X} \to \mathbb{R}$, and under the bounded cost per stage assumption (6.3), the iterates $T^k J$ converge uniformly to J^*:*

$$\lim_{k \to \infty} \|T^k J - J^*\|_\infty \to 0.$$

Proof. For every K, x_0 and policy $\pi = \{\mu_0, \mu_1, \ldots\}$, we have

$$J^\pi_\alpha(x_0) = E\left[\sum_{k=0}^{K-1} \alpha^k c(x_k, \mu_k(x_k), w_k)\right] + \lim_{N \to \infty} E\left[\sum_{k=K}^{N} \alpha^k c(x_k, \mu_k(x_k), w_k)\right].$$

Under the bounded cost per stage assumption (6.3), we have

$$\left| \lim_{N \to \infty} E\left[\sum_{k=K}^{N} \alpha^k c(x_k, \mu_k(x_k), w_k)\right] \right| \leq \frac{\alpha^K M}{1 - \alpha}.$$

Hence

$$J_\alpha^\pi(x_0) - \frac{\alpha^K M}{1-\alpha} - \alpha^K \|J\|_\infty \leq E\left[\sum_{k=0}^{K-1} \alpha^k c(x_k, \mu_k(x_k), w_k) + \alpha^K J(x_K)\right]$$

$$\leq J_\alpha^\pi(x_0) + \frac{\alpha^K M}{1-\alpha} + \alpha^K \|J\|_\infty.$$

Taking now the infimum over policies, we get for all x_0 and K (rigorously, consider the infimum for one inequality at a time, from left to right, to get this):

$$J^*(x_0) - \frac{\alpha^K M}{1-\alpha} - \alpha^K \|J\|_\infty \leq (T^K J)(x_0) \leq J^*(x_0) + \frac{\alpha^K M}{1-\alpha} + \alpha^K \|J\|_\infty.$$

The result follows by letting $K \to \infty$. □

6.4 Contraction Mappings

The next set of results, concerning Bellman's equation, will follow directly from the properties of the DP operator, in particular from the contraction property. We recall some of the necessary background on contraction mappings in this section. Let (X, d) be a complete metric space. A map $F : X \to X$ is said to be a *contraction* if there is a constant $\lambda < 1$ such that $d(F(x), F(y)) \leq \lambda d(x, y)$ for all $x, y \in X$. Sometimes, λ is called the *modulus of contraction.*

Exercise 14. Show that any contraction is uniformly continuous (in fact, it satisfies the definition of Lipschitz continuity).

Theorem 6.4.1 (Banach fixed point theorem or contraction principle)**.** *Let X be a nonempty* complete *metric space. Every contraction $F : X \to X$ has a unique fixed point x^*, i.e., a point in X such that $F(x^*) = x^*$. Furthermore if $x \in X$ then $d(F^k x, x^*) \leq \lambda^k d(x, x^*)$.*

Proof. To show uniqueness, suppose x, x' are both fixed points. Then $d(x, x') = d(F(x), F(x')) \leq \lambda d(x, x')$, which only makes sense if $x = x'$. The second part of the theorem tells us how to prove the existence of a fixed point and a constructive method for computing it. Let $x_0 \in X$ and consider the sequence of iterates $x_{n+1} = F(x_n), n \geq 0$. We have $d(x_i, x_{i+1}) = d(F(x_{i-1}), F(x_i)) \leq d(x_{i-1}, x_i)$ and so by immediate induction

$$d(x_i, x_{i+1}) \leq \lambda^i d(x_0, x_1).$$

Then if $j \geq i \geq k$,

$$\begin{aligned} d(x_i, x_j) &\leq d(x_i, x_{i+1}) + d(x_{i+1}, x_{i+2}) + \ldots + d(x_{j-1}, x_j) \\ &\leq (\lambda^i + \ldots + \lambda^{j-1}) d(x_0, x_1) \\ &\leq \lambda^i \left(\sum_{n=0}^{\infty} \lambda^n \right) d(x_0, x_1) \\ &\leq \frac{\lambda^k}{1-\lambda} d(x_0, x_1). \end{aligned}$$

Hence $\{x_n\}_{n \geq 0}$ is a Cauchy sequence, and therefore converges to a limit $\hat{x} \in X$ since X is complete. Since F is continuous,

$$F(\hat{x}) = F(\lim_{n \to \infty} x_n) = \lim_{n \to \infty} F(x_n) = \lim_{n \to \infty} x_{n+1} = \hat{x}.$$

So $\hat{x}$ is the desired fixed point. For the last part, just note that

$$d(F^k x, x^*) = d(F^k x, F^k x^*) \leq \lambda^k d(x, x^*).$$

□

Remark. The Banach fixed point theorem is an elementary result that forms the basis of a number of important applications, so you should remember it. You probably used it before to show the existence of the solutions of ODEs, or to prove the inverse function theorem in calculus.

Since we know from lemma (6.3.3) that T is a contraction on $(B(\mathsf{X}, \mathbb{R}), \| \cdot \|_\infty)$, we can apply the contraction principle if we know that $B(\mathsf{X}, \mathbb{R})$ is complete for the sup-norm $\| \cdot \|_{\infty}$. This is in fact a classical result in analysis. For our purposes later on, we will consider the slightly more general *weighted* sup-norm. Let $v : \mathsf{X} \to \mathbb{R}$ be a function such that $v(x) > 0$ for all $x \in \mathsf{X}$. Then we define

$$\|f\|_{v,\infty} = \sup_{x \in \mathsf{X}} \left| \frac{f(x)}{v(x)} \right|,$$

for all functions $f : \mathsf{X} \to \mathbb{R}$. It is easy to show that this defines a norm. Then denote (the vector space)

$$B_v(\mathsf{X}, \mathbb{R}) = \Big\{ f : \mathsf{X} \to \mathbb{R} \Big| \|f\|_{v,\infty} < \infty \Big\}.$$

Lemma 6.4.2. *The vector space $B_v(\mathsf{X}, \mathbb{R})$ is complete for the norm $\| \cdot \|_{v,\infty}$.*

Exercise 15 (optional)**.** Prove lemma 6.4.2. You can assume $v(x) = 1$ for all x (hint: start by showing pointwise convergence using the fact that $\mathbb{R}$ is complete).

6.5 Bellman's Equation for Discounted Cost Problems with Bounded Cost Per Stage

From lemma 6.3.3, 6.3.4, and the results in section 6.4, the next theorem is just a paraphrase of the contraction principle.

Theorem 6.5.1. *Assume that $0 \leq \alpha < 1$ and that the bounded cost-per-stage assumption (6.3) holds. The optimal cost function J^* is the unique fixed point in $B(\mathsf{X}, \mathbb{R})$ of the DP operator T, i.e., it satisfies* Bellman's equation

$$J^* = TJ^* \tag{6.7}$$

Moreover for any bounded function $J : \mathsf{X} \to \mathbb{R}$, the sequence $T^k J$ converges uniformly (and linearly with rate α) to the optimal cost J^*

$$\|J^* - T^k J\|_\infty \leq \alpha^k \|J^* - J\|_\infty.$$

Similarly the cost J_μ of the stationary policy $\{\mu, \mu, \ldots\}$ is the uniques fixed point of T_μ in $B(\mathsf{X}, \mathbb{R})$, and so it satisfies the equation

$$J_\mu = T_\mu J_\mu,$$

and we have, for any $J : \mathsf{X} \to \mathbb{R}$

$$\|J_\mu - T_\mu^k J\|_\infty \leq \alpha^k \|J_\mu - J\|_\infty.$$

Finally, a stationary policy μ is optimal ($J_\mu = J^*$) if and only if $\mu(x)$ attains the minimum in Bellman's equation (6.7) for each $x \in \mathsf{X}$, i.e.,

$$TJ^* = T_\mu J^*. \tag{6.8}$$

Proof. We have proved in lemma 6.3.4 that the iterates $T^k J$ converge to the optimal cost J^*. They also converge to the unique fixed point of T by the Banach fixed point theorem, so J^* is the unique solution of the fixed point equation 6.7. It only remains to prove the characterization of the optimal policy. If μ is optimal, then $J_\mu = J^*$ and so

$$T_\mu J^* = T_\mu J_\mu = J_\mu = J^* = TJ^*.$$

Conversely if (6.8) is satisfied, then by Bellman's equation

$$T_\mu J^* = TJ^* = J^*,$$

so J^* is a fixed point of T_μ, and by unicity we must have $J^* = J_\mu$. So μ is optimal.

□

We have seen earlier the value iteration algorithm, which starts with any bounded function J, for example $J \equiv 0$, and compute the iterates $T^k J$. This produces a sequence of functions that converges uniformly to the value function J^*. Moreover, if we know J^*, then we can obtain the optimal policy μ^* from (6.8): in state x, $\mu^*(x)$ attains the minimum in Bellman's equation. More explicitely, Bellman's equation for a discounted cost problem is

$$J^*(x) = \min_{u \in \mathsf{U}(x)} E\left[c(x,u,w) + \alpha J^*(f(x,u,w))\right].$$

If X is finite, this is a system of nonlinear equations, with unknowns the values $\{J(x)\}_{x \in \mathsf{X}}$. This system can also be written, for a controlled Markov chain model,

$$J^*(x) = \min_{u \in \mathsf{U}(x)} \left[\sum_{y \in \mathsf{X}} p_{xy}(u) \Big[c(x,u,y) + \alpha J^*(y) \Big] \right].$$

If for all x this minimum is attained at $\mu^*(x)$, then μ^* is a stationary optimal policy. Finally for J_μ we would write

$$J_\mu(x) = \left[\sum_{y \in \mathsf{X}} p_{xy}(\mu(x)) \Big[c(x,\mu(x),y) + \alpha J_\mu(y) \Big] \right].$$

Using the iterations $T_\mu^k J$ to compute the cost J_μ of a stationary policy μ is called *policy evaluation*.