

Chapter 0

Discrete Time Dynamic Programming

0.1 The Finite Horizon Case

Time is discrete and indexed by $t = 0, 1, \dots, T$, where $T < \infty$. An individual is interested in maximizing an objective function given by

$$E_0 \sum_{t=0}^T \beta^t u(x_t, a_t), \quad (0.1)$$

where the instantaneous return function (or utility function) u depends on a *state variable* x_t and a *control variable* a_t at each date. We assume that $x_t \in X \subset \mathcal{R}^m$, $a_t \in A \subset \mathcal{R}^n$ for all t , $\beta \in (0, 1)$, and E_0 denotes the expectation conditional on information available at date 0. The nature of the uncertainty in the environment will be made explicit below.

There are several constraints faced by the individual. At each date the control variable is constrained to belong to a set that may depend on the state: $a_t \in \Gamma(x_t)$ for all t . The initial value of the state x_0 is given by nature. Future states evolve according to the *transition equation*

$$x_{t+1} = f(x_t, a_t, \varepsilon_t), \quad (0.2)$$

where ε_t is a random variable, unobservable at date t , described by a cumulative distribution function that may depend on the state and action but does not depend on t

$$F(\varepsilon \mid x, a) = \text{Prob}(\varepsilon_t \leq \varepsilon \mid x_t = x, a_t = a). \quad (0.3)$$

Given any sequence of controls $\{a_t\}$, we can construct the probability distribution of future states conditional on x_0 from (0.2) and (0.3); the expectation E_0 in (0.1) is with respect to this distribution.¹

Notice that the environment here is *stationary*, in the sense that u , Γ , f and F do not depend on t . It would be a straightforward generalization to allow these objects to vary with time, but we adopt a stationary environment because it yields time-invariant decision rules in the infinite horizon case studied below. In any case, (x_t, a_t) contains all of the information available at date t that is relevant for the probability distribution of future events. This, together with the additive separability of the objective function in (0.1), implies that the control at date t will depend only on the current state, $a_t = \alpha_t(x_t)$, where α_t is referred to as the *decision rule*.

A *policy* (of length T) is defined to be a sequence of decision rules, $\pi_T = (\alpha_0, \alpha_1, \dots, \alpha_T)$, where $\alpha_t : X \rightarrow A$ for all t . The set of feasible policies is

$$\Pi_T = \{\pi_T = (\alpha_0, \alpha_1, \dots, \alpha_T) : \alpha_t(x) \in \Gamma(x) \quad \forall x, t\}. \quad (0.4)$$

A policy is stationary if it does not depend upon time: $\alpha_t(x) \equiv \alpha(x)$. Any given policy generates a stochastic law of motion for the state,

$$x_{t+1} = f[x_t, \alpha_t(x_t), \varepsilon_t],$$

which will be stationary if α_t is stationary.

¹Alternatively, the uncertainty can be described by starting with a Markov probability transition function,

$$Q(x, a, \tilde{X}) = \text{Pr}(x_{t+1} \in \tilde{X} \mid x_t = x, a_t = a),$$

for $\tilde{X} \subset X$, as in Stokey, Lucas and Prescott (1989). One can always construct such a Q from our f and F ; see Stokey et al. (especially Theorem 8.9) for a rigorous treatment.

Define the *value* of following policy π_T when there are T periods left to go and the current state is x_0 by

$$W_T(x_0, \pi_T) = E_0 \sum_{t=0}^T \beta^t u[x_t, \alpha_t(x_t)], \quad (0.5)$$

where it is understood that x_t evolves according to (0.2). The individual's problem is to choose a $\pi_T \in \Pi_T$ to maximize $W_T(x_0, \pi_T)$. If we assume that the constraint correspondence $\Gamma(x)$ is non-empty, compact and continuous, $u(x, a)$ is continuous and bounded, and $f(x, a, \varepsilon)$ is continuous, then there exists a solution to the problem, $\pi_T^* = (a_0^*, a_1^*, \dots, a_T^*)$, called an *optimal policy*. Furthermore, under these conditions, the *optimal value function*

$$V_T(x) = W_T(x, \pi_T^*) \quad (0.6)$$

exists, and is bounded and continuous in x .²

The law of iterated expectations says that $E_0(\cdot) = E_0[E_1(\cdot)]$, where E_t is the expectation conditional on x_t . Hence, we can write the optimal value function as

$$V_T(x_0) = \max_{\pi_T \in \Pi_T} E_0 \left\{ u(x_0, a_0) + E_1 \sum_{t=1}^T \beta^t u(x_t, a_t) \right\}.$$

Since decisions at future dates $t \geq 1$ do not affect the instantaneous return at $t = 0$, we can cascade the maximization operator and write

$$V_T(x_0) = \max_{a_0 \in \Gamma(x_0)} E_0 \left\{ u(x_0, a_0) + \max_{\pi_{T-1} \in \Pi_{T-1}} E_1 \sum_{t=1}^T \beta^t u(x_t, a_t) \right\},$$

²Assuming that f is continuous guarantees that the stochastic structure satisfies the *Feller property*, which is that the function

$$E[\varphi(x_{t+1}) \mid x_t = x, a_t = a] = \int \varphi[f(x, a, \varepsilon)] dF(\varepsilon \mid x, a)$$

is bounded and continuous in (x, a) for every bounded and continuous real valued function φ (see Stokey et al. 1989, exercise 8.10). Given the Feller property, the existence of an optimal policy and the continuity of the value function, which follows from the Theorem of the Maximum, can be established as in Stokey et al. (1989).

where Π_{T-1} is the set of feasible policies with $T - 1$ periods left to go. Let $V_{T-1}(x_1)$ be the value function with $T - 1$ periods to go starting from $x_1 = f(x_0, a_0, \varepsilon_0)$. Then we can write

$$V_T(x_0) = \max_{a_0 \in \Gamma(x_0)} E_0 \{u(x_0, a_0) + \beta V_{T-1}[f(x_0, a_0, \varepsilon_0)]\}.$$

If we omit the time subscripts on x , a , and ε (because they are all evaluated at the same date) and rewrite the previous expression for any number of periods $S \in \{1, 2, \dots, T\}$ left to go, we arrive at *Bellman's equation*:

$$V_S(x) = \max_{a \in \Gamma(x)} \{u(x, a) + \beta EV_{S-1}[f(x, a, \varepsilon)]\}. \quad (0.7)$$

The expectation in (0.7) is with respect to ε and is understood to be conditional on x and a ; that is,

$$EV_{S-1}[f(x, a, \varepsilon)] = \int V_{S-1}[f(x, a, \varepsilon)] dF(\varepsilon | x, a).$$

Bellman's equation expresses the choice of a sequence of decision rules as a sequence of choices for the control variable, which simplifies the problem considerably.

In particular, it leads to the following solution procedure known as the *DP algorithm* (see, e.g., Bertsekas 1976 for an extensive discussion). Start at the final period, with $S = 0$ periods left to go, and construct

$$V_0(x) = \max_{a \in \Gamma(x)} u(x, a)$$

by solving the maximization problem for a . Since the solution will generally depend on the value of x , we write $a = \eta_0(x)$, where the subscript indicates that there are 0 periods left to go. Then, work backwards, solving the maximization problem in (0.7) with S periods to go for each $S = 1, 2, \dots, T$, constructing $V_S(x)$ and recording the decision rule $\eta_S(x)$ at each step. We can then construct a policy by setting $\alpha_t(x) = \eta_{T-t}(x)$, for $t = 0, 1, \dots, T$. The policy $\pi = (\alpha_0, \alpha_1, \dots, \alpha_T)$ is optimal.

0.2 The Infinite Horizon Case

When the horizon is infinite, we cannot proceed directly with the DP algorithm, since there is no last period in which to start. However, it seems natural to regard the infinite horizon problem as the limit of a sequence of finite problems as $T \rightarrow \infty$. Conversely, when T is large but finite, it would seem convenient to study the infinite horizon as an approximation to the finite problem. The justification for either approach requires the demonstration that there exists a unique limit of the sequence of finite horizon optimal policies and this limit is the optimal policy for the infinite horizon problem.

In the infinite horizon case, because the environment is stationary, the problem is the same at each point in time. Therefore, *if* a solution to the problem exists, the value function $V(x)$ and decision rule $a = \alpha(x)$ will be the same at each point in time, and will satisfy

$$V(x) = \max_{a \in \Gamma(x)} \{u(x, a) + \beta EV[f(x, a, \varepsilon)]\} \quad (0.8)$$

where, as in (0.7), the expectation is with respect to ε and is understood to be conditional on x and a . Equation (0.8) is a *functional equation* – that is, an equation whose unknown is a function, in this case the function V . We are interested in knowing if it has a solution.

It is useful to think of (0.8) in the following way. Let \mathcal{C} be the set of continuous and bounded real-valued functions on X (note that continuous functions are automatically bounded if X is compact, but not if X is arbitrary; hence, the condition that the functions are bounded is not redundant). Then, define a mapping $T : \mathcal{C} \rightarrow \mathcal{C}$ as follows:

$$T\varphi = \max_{a \in \Gamma(x)} \{u(x, a) + \beta E\varphi[f(x, a, \varepsilon)]\}. \quad (0.9)$$

Thus, T takes one function φ and maps it into another function $T\varphi$.³ A solution to (0.8) is a fixed point of T ; that is, the function V in Bellman's

³The fact that $T\varphi \in \mathcal{C}$ for any $\varphi \in \mathcal{C}$ follows from the Feller property discussed earlier (the proof is a straightforward generalization of Lemmas 9.5 and 12.14 in Stokey et al. 1989)

equation satisfies $V = TV$. To look for fixed points of T , it is helpful to have a few mathematical tools at our disposal.

A metric space is a set \mathcal{M} together with a metric (or distance function) $d : \mathcal{M} \times \mathcal{M} \rightarrow \mathcal{R}$ with the following properties: for all φ, ψ , and ζ in \mathcal{M} ,

- a) $d(\varphi, \psi) \geq 0$;
- b) $d(\varphi, \psi) = d(\psi, \varphi)$;
- c) $d(\varphi, \psi) + d(\psi, \zeta) \leq d(\varphi, \zeta)$.

For example, Euclidean space \mathcal{R}^n together with the usual notion of distance,

$$d(\varphi, \psi) = \left[(\varphi_1 - \psi_1)^2 + \dots + (\varphi_n - \psi_n)^2 \right]^{\frac{1}{2}},$$

is a metric space. So is the set \mathcal{C} of continuous and bounded real-valued functions on X together with $d(\varphi, \psi) = \|\varphi - \psi\|$, where $\|\varphi\| = \sup |\varphi(x)|$ is the sup norm.

The notion of a metric space allows us to consider convergence. A sequence $\{\varphi_n\}$ in \mathcal{M} is said to converge to φ in \mathcal{M} if: for all $\delta > 0$ there exists $N = N(\delta) > 0$ such that $d(\varphi_n, \varphi) < \delta$ if $n > N$. A related concept is that of a Cauchy sequence, which is a sequence $\{\varphi_n\}$ in \mathcal{M} with the following property: for all $\delta > 0$ there exists $N = N(\delta)$ such that $d(\varphi_n, \varphi_m) < \delta$ if $n, m > N$. It is easily verified that if $\{\varphi_n\}$ converges then it is a Cauchy sequence, but the converse is not true in general (see below). In some metric spaces, however, all Cauchy sequences do converge, and we call such metric spaces *complete*. One can show that the metric space defined by \mathcal{C} together with the sup norm is complete (see Stokey et al. 1989, Theorem 3.1).

Now consider a metric space (\mathcal{M}, d) and a mapping $\Psi : \mathcal{M} \rightarrow \mathcal{M}$ that satisfies $d(\Psi\varphi, \Psi\psi) \leq \beta \cdot d(\varphi, \psi)$ for all $\varphi, \psi \in \mathcal{M}$, for some $\beta < 1$; such a mapping is called a *contraction* (of modulus β). The name comes from the fact that Ψ contracts points, in the sense that $\Psi\varphi$ and $\Psi\psi$ are closer together than φ and ψ . Blackwell (1965) provides us with the following result that can often be easily used to show that something is a contraction. Suppose that \mathcal{M} is the set of bounded real-valued functions on $X \subset \mathcal{R}^m$ and the metric is

given by the sup norm; if $\Psi: \mathcal{M} \rightarrow \mathcal{M}$ satisfies the following two conditions,

- a) $\varphi \leq \psi \Rightarrow \Psi\varphi \leq \Psi\psi \quad \forall \varphi, \psi \in \mathcal{C}$
- b) $\Psi(a + \varphi) \leq a\beta + \Psi\varphi \quad \forall a > 0, \varphi \in \mathcal{C}$,

for some $\beta < 1$, then Ψ is a contraction. See Stokey et al. (1989), Theorem 3.3.

Blackwell's conditions can be used to prove that the mapping T defined by (0.9) is a contraction with modulus β on the metric space given by \mathcal{C} together with the sup norm. We leave this as an exercise, and instead provide a direct proof. Let \bar{a} solve the maximization problem in the definition of T ; then

$$\begin{aligned}
 T\varphi &= u(x, \bar{a}) + \beta E\varphi[f(x, \bar{a}, \varepsilon)] \\
 &= u(x, \bar{a}) + \beta E\psi[f(x, \bar{a}, \varepsilon)] \\
 &\quad + \beta E\{\varphi[f(x, \bar{a}, \varepsilon)] - \psi[f(x, \bar{a}, \varepsilon)]\} \\
 &\leq \max_{a \in \Gamma(x)} \{u(x, a) + \beta E\psi[f(x, a, \varepsilon)]\} + \beta\|\varphi - \psi\| \\
 &= T\psi + \beta\|\varphi - \psi\|.
 \end{aligned}$$

This implies $T\varphi - T\psi \leq \beta\|\varphi - \psi\|$. Reversing the roles of φ and ψ implies $T\psi - T\varphi \leq \beta\|\varphi - \psi\|$. Hence, $\|T\varphi - T\psi\| \leq \beta\|\varphi - \psi\|$, and T is a contraction.

The Contraction Mapping Theorem (or Banach Fixed Point Theorem) tells us that, if (\mathcal{M}, d) is a complete metric space and $\Psi : \mathcal{M} \rightarrow \mathcal{M}$ is a contraction then two things are true: first, Ψ has a unique fixed point (i.e., there is a unique $\varphi^* \in \mathcal{M}$ such that $\Psi\varphi^* = \varphi^*$); second, the sequence defined by $\varphi_{n+1} = \Psi\varphi_n$ will converge to the fixed point φ^* for any initial condition φ_0 (see Stokey et al. Theorem 3.2).⁴

⁴The reason that the metric space has to be complete can be seen from the following example. The mapping from the open interval $(0, 1)$ into itself defined by $\Psi(x) = bx$ is a contraction if and only if $0 < b < 1$. Now for any $x_0 \in (0, 1)$, the sequence

$$x_n = bx_{n-1} = b^n x_0$$

is a Cauchy sequence but does not converge in $(0, 1)$. It converges, of course, to 0; but 0 does not belong to the interval in question. Complete metric spaces are those for which Cauchy sequences converge, which avoids the sort of problem illustrated by this example.

Because the metric space defined by \mathcal{C} together with the sup norm is complete the mapping T defined in (0.9) has a unique fixed point in \mathcal{C} , and hence there exists a unique $V \in \mathcal{C}$ satisfying Bellman's equation. Furthermore, the sequence defined by $V_0 = 0$ and $V_{n+1} = TV_n$ converges to V in the sup norm (i.e., *uniformly*). Because V_n can be regarded as the value function constructed using the DP algorithm for a finite problem with horizon n , V is indeed the limit of the sequence of finite horizon value functions from progressively longer problems.

Moreover, if we assume there is a finite number of periods to go, n , we can apply the DP algorithm to find $a = \eta_n(x)$, exactly as described above. If η_n is a single-valued function (at least for sufficiently large n), then one can show that η_n converges pointwise to an optimal stationary policy for the infinite horizon problem, α , and if the state variable x is restricted to a compact set the convergence is uniform (see Stokey et al. 1989, Theorem 3.8). Thus, the infinite horizon dynamic programming problem is in the relevant sense an approximation to a long finite horizon problem, and we can find the value function V and decision rule α for the infinite horizon problem by iterating on Bellman's equation.

We now describe how some properties of value functions can be derived from exploiting the contraction property. First, note that a closed subset of a complete metric space is also a complete metric space. Hence, if we can show that T maps a closed set $\mathcal{D} \subset \mathcal{C}$ into itself then we can apply the contraction mapping theorem to the restriction of T to \mathcal{D} . For example, let \mathcal{D} be the set of increasing functions, which is a closed subset of \mathcal{C} . Under suitable restrictions on the primitive functions u , f , F and Γ (see below), one can show that $T\varphi$ is increasing whenever φ is increasing; therefore, the fixed point V must be increasing. A similar argument can also be used to show that under suitable restrictions $T\varphi$ is concave whenever φ is concave; therefore, V must be concave.⁵

⁵This style of argument cannot be used to show that V is differentiable, because the set of differentiable functions is not closed in \mathcal{C} . But different styles of argument can be

The above arguments could not be used to show that the value function *strictly* increasing, as the set of strictly increasing functions is not closed. However, a two-step argument can be used to show that the value function is strictly increasing. Suppose that under certain conditions T maps increasing functions into strictly increasing functions. Then we can first establish that V is increasing by the above argument, and then observe that it must be strictly increasing because T maps the increasing function V into the strictly increasing function TV , and $TV = V$. A similar two-step argument can be used to show under certain conditions the value function is strictly concave, which is a useful result in some applications because it could be used to show that there is a unique solution to the maximization problem in Bellman's equation, which would imply that the decision rule is a single-valued function.

To illustrate, suppose that the constraint set $\Gamma(x)$ is convex for all x , and that u and f are increasing in x . Then for any increasing function φ , for any two points x_1 and $x_2 > x_1$, and for any a_0 ,

$$\begin{aligned} T\varphi(x_2) &\geq u(x_2, a_0) + \beta E\varphi[f(x_2, a_0, \varepsilon)] \\ &\geq u(x_1, a_0) + \beta E\varphi[f(x_1, a_0, \varepsilon)], \end{aligned}$$

where T is the mapping defined in (0.9). Maximizing the right hand side over a_0 , we have $T\varphi(x_2) \geq T\varphi(x_1)$. Hence, $T\varphi$ is increasing, and we conclude T takes increasing functions into increasing functions. Thus, the value function $V = \lim T^N \varphi$ is increasing. We leave as an exercise the verification of the result that if u and f are concave in (x, a) as well as increasing in x , then V is also concave, also the generalization of both results to show V is strictly increasing or concave.

We now introduce the notion of *unimprovability*. Let the value of following an arbitrary stationary policy α starting from state x be $W(x, \alpha)$; this is a stationary version of the function $W_T(x, \pi_T)$ defined in (0.5). Now consider

used to prove V is differentiable (see Stokey et al. 1989), and even twice differentiable (see Santos 1991 and Araujo 1991), at least in an important class of dynamic programming problems.

the following mapping

$$\hat{T}W(x, \alpha) = \max_{a \in \Gamma(x)} \{u(x, a) + \beta EW[f(x, a, \varepsilon), \alpha]\}, \quad (0.10)$$

which should be compared with (0.9). Notice that $\hat{T}W(x, \alpha)$ is the value of choosing the best value of a now and reverting to the stationary decision rule α next period. If the solution to the maximization problem in (0.10) is $a = \alpha(x)$ for all x , then we say that the decision rule α is unimprovable in a single step, or, for short, simply *unimprovable*.

It is obvious that an optimal policy α^* is unimprovable; this merely says that if you are using the best policy then you cannot improve your payoff by a one-shot deviation. What is perhaps less obvious, although still true, is that any unimprovable policy is optimal; this says that if you cannot improve your payoff by a one-shot deviation then you cannot improve your payoff with any deviation. In other words, given an optimal policy α^* , we have $\hat{T}W(x, \alpha^*) = W(x, \alpha^*) = V(x)$.⁶

Unimprovability is extremely useful in many applications, for in order to verify that a candidate policy is optimal it suffices to show that there is no incentive to deviate from it at any single point in time. For example, suppose that consider the policy $\alpha(x) = a_0$ for all x . To check that this is optimal, it suffices to check that an agent always choosing a_0 cannot do better by deviating at a single point in time from this policy and choosing $a \neq a_0$. Of course, we have to check this for every state x at that point in time, but we do not have to check that the agent cannot do better by deviating at two consecutive dates, at all future dates, at all even dates, and so on.

Finally, we show how the basic job search model fits neatly into the above framework. Let the state variable $x \in X$ be the current wage offer, and let the control variable a be constrained to the set $\Gamma(x) = \{0, 1\}$, where $a = 1$ indicates that the offer is accepted while $a = 0$ indicates that it is

⁶The result that an unimprovable policy is optimal requires $u(x, a)$ bounded from below; see Kreps (1990) for a very readable discussion of this result, together with an example showing what can go wrong when $u(x, a)$ is not bounded from below.

rejected. The instantaneous return is $u(x, a) = aU(x) + (1 - a)U(b)$, where $U : X \rightarrow \mathcal{R}$ is bounded and differentiable with $U' > 0$ and $U'' \leq 0$, and b is a constant representing unemployment income. The transition equation $x_{t+1} = f(x_t, a_t, \varepsilon_t)$ is defined as follows: $x_{t+1} = x_t$ if $a = 1$ and $x_{t+1} = \varepsilon_t$ if $a = 0$, where ε_t is drawn from a distribution F that in this case does not depend on (x, a) . Thus, accepted offers are retained forever while rejected offers are lost and replaced by random new offers next period.

The value function is the unique $V \in \mathcal{C}$ that solves Bellman's equation,

$$V(x) = \max_{a \in \{0,1\}} \{aU(x) + (1 - a)U(b) + \beta EV[f(x, a, \varepsilon)]\},$$

where $EV[f(x, 1, \varepsilon)] = V(x)$ and $EV[f(x, 0, \varepsilon)] = \int V(\varepsilon)dF(\varepsilon)$. It is obviously equivalent to write Bellman's equation in this case as

$$V(x) = \max \left\{ U(x) + \beta V(x), U(b) + \beta \int V(\varepsilon)dF(\varepsilon) \right\}.$$

An optimal policy is to set $a = \alpha(x) = 1$ if and only if

$$U(x) + \beta V(x) \geq U(b) + \beta \int V(\varepsilon)dF(\varepsilon).$$