Chapter 3

Difference Equations

Difference equations are the discrete-time analog to differential equations. For a differential equation of the form

$$\dot{y}(t) = f\big(x(t), y(t)\big) \,,$$

the discrete-time analog is

$$y_t - y_{t-1} = f\big(x_{t-1}, y_{t-1}\big)$$

or

$$y_t = \tilde{f}\big(x_{t-1}, y_{t-1}\big) \,.$$

Difference equations are valuable alternatives to differential equations for a number of reasons:

- They lend themselves readily to econometric analysis because difference equations can be written to match data frequencies. Empirically-implementing continuous time models with data only observed at discrete intervals requires some fairly advanced econometrics (e.g. Bergstrom [1990]; in practice, however, researchers just ignore the complications created in the mapping from continuous time modeling to discrete time empirical work).

- They have been the main mode of analysis for the once-trendy fields of chaos and complexity.

- They lend themselves more readily to the incorporation of rational expectations.

- They are more easily extended then differential equations to deal with stochastic models. One notable exception is the Poisson-family of models of rare events, which is best analyzed in continuous time.

- They are central to the analysis of many models of dynamic programming (an approach to dynamic optimization that we study later in the course).

In these notes, as in other sections, we provide only a basic introduction to difference equations. A significant fraction of the material is devoted to models of rational expectations under uncertainty. Although such models have been in the core of macroeconomic dynamics for thirty years, they are in my opinion still underutilized in microeconomic dynamics.

1. Deterministic Difference Equations

We consider here first-order linear difference equations with constant coefficients:

$$c_1 y_t + c_0 y_{t-1} = g_t \,,$$

where $c_1 \neq 0$ and $c_2 \neq 0$ (otherwise it is not a difference equation).

The Homogeneous Equation

Define $b{=}c_0/c_1$ and write the homogeneous equation in which $g(t){=}0$ for all t:

$$y_t + b y_{t-1} = 0 \,.$$

Let y_0 be the initial condition. Then it is easy to verify that the solution is

$$y_t = y_0(-b)^t \,.$$

The analog to the solution of a differential equation of the same type should be immediately apparent. To verify that this is the solution, substitute the solution into the difference equation:

$$y_t + b y_{t-1} = 0 \quad \Rightarrow \quad y_0(-b)^t + b y_0(-b)^{t-1} = 0 \quad \Rightarrow \quad y_0(-b)^t - y_0(-b)^t = 0 \quad \Rightarrow \quad 0 = 0 \,.$$

Of course, if y_0 is not known we have only the *form* of the solution.

Note that

- For $|b| < 1$, $\lim_{t\to\infty} |y_t| = 0$ and $|y_{t+1}| < |y_t|$.
- For $b<0$ the path is monotonic while for $b>0$ the path oscillates.

The Inhomogeneous Equation

Consider now the case $c_1 y_t + c_0 y_{t-1} = g_t$. For most common functions, this can be solved by the *method of undetermined coefficients*. The method involves guessing a functional form with unknown coefficients, and then verifying the guess and obtaining the value of the unknown coefficients . We give some examples here.

- ***g*$_t$ is a constant,** $c_1 y_t + c_0 y_{t-1} = a$. Try $y_t=\mu$ for some unknown μ. If this guess is correct then $c_1\mu + c_0\mu = a$ or $\mu = a/(c_0 + c_1)$. Moreover, we can verify that the guess is correct because we see that the equation implies that μ does turn out to be a constant.[1] This solution does not work if $c_0+c_1=0$. In this case, guess $y_t=\mu t$. Then, it is easy to verify that the guess is correct and $\mu=a/c_0$ or, equivalently, $-\mu=a/c_1$.

- ***g*$_t$ is exponential**, $c_1 y_t + c_0 y_{t-1} = k\beta^t$. Try $y_t=\mu\beta^t$ for some unknown μ. If this guess is correct then $c_1\mu\beta^t + c_0\mu\beta^{t-1} = k\beta^t$, which is true if and only if $c_1\mu\beta + c_0\mu - k\beta = 0$ [i.e., if $\mu = k\beta/(c_0 + \beta c_1)$]. This solution fails if $\beta c_0+c_1=0$. In this case, guess $y_t=\mu t\beta^t$.

---

[1] Imagine we have an equation of the form $c_1 y_t + c_0 y_{t-1} = at$, and now we try $y_t = \mu$ for some unknown μ. If this guess is correct then $c_1\mu + c_0\mu = at$. But this implies that $\mu = at/(c_0 + c_1)$. As μ depends on t, then it clearly is not a constant; thus our initial guess must have been wrong and we should try another. Now return to the original equation $c_1 y_t + c_0 y_{t-1} = a$, and assume we guess $y_t = \mu t$ for some unknown μ. If this guess is correct then $c_1\mu t + c_0\mu(t-1) = a$. But this implies that $\mu = a/(c_0(t-1) + c_1 t)$. Yet again, μ depends on t (unless $c_0 + c_1 = 0$) and our initial guess must have been wrong. These simple examples show that each time the wrong functional form is guessed you will end up with an inconsistency that sends you back for another guess.

NOTE: *If the function you try does not work for some values of the coefficients, try the same function multiplied by t. This is a quite general prescription regardless of the form of* g_t.

- **g_t is a polynomial of degree m.** Guess $y_t = a_0 + a_1 t + \cdots + a_m t^m$, with unknown coefficients a_i.

- **g_t is a trigonometric function of the sine-cosine type** (if you have a model of this sort, you are probably not doing economic modeling!). If $g_t = a\cos\omega t + b\sin\omega t$, guess a solution of the form $y_t = \alpha\cos\omega t + \beta\sin\omega t$, with unknown coefficients α and β.

We will take more time here to think about solution methods for the case where g_t is an arbitrary, unspecified function. Obviously, one cannot in this case apply the method of undetermined coefficients. To tackle the case of g_t unspecified, we will rely on *operational methods*.

DEFINITION: (lag and forward operators). *Let L and F denote respectively, a lag operator and a forward operator. L and F have the following properties.*

$$Ly_t \equiv F^{-1}y_t \equiv y_{t-1}\,, \qquad L^n y_t \equiv F^{-n} \equiv y_{t-n}\,,$$
$$L^{-1}y_t \equiv Fy_t \equiv y_{t+1}\,, \qquad L^{-n}y_t \equiv F^n y_t \equiv y_{t+n}\,.$$

It should be apparent from the definition that the operators L and F are just a different way to write the time index on variables. What makes this different notation valuable is that L and F can be manipulated just like any other algebraic quantity. For example, the equation $y_t + by_{t-2} = ax_{t-1}$ can be written as $y_t + bL^2 y_t = aLx_t$. We can then divide throughout by L^2 and use the fact that $L^{-2} = F^2$ to get $F^2 y_t + by_t = aFx_t$, or $y_{t+2} + by_t = ax_{t+1}$. Obviously, this is a simple updating of the equation. However, in many instances, updating is not so straightforward and the operators turn out to be useful.

Before providing some examples of nontrivial uses of the operators, let us note two extremely valuable series expansions:

$$(1-\alpha L)^{-1} = \sum_{i=0}^{\infty} \alpha^i L^i \,, \qquad \text{for } |\alpha| < 1\,, \tag{1.1}$$

$$(1-\alpha L)^{-1} = -\sum_{i=1}^{\infty} \alpha^{-i} L^{-i} \qquad \text{for } |\alpha| > 1\,. \tag{1.2}$$

Please note the difference in the lower limits of the summation in the two series expansion. The reason we need two expansions is that if $|\alpha|>1$ the infinite series in (1.1) diverges (i.e. each successive term in α becomes larger and larger). In contrast, the infinite series in (1.2) diverges if $|\alpha|<1$.

Heuristic proofs of these expansions are easy. For (1.1),

$$\begin{aligned}
(1-\alpha L)^{-1}(1-\alpha L) = 1\,, \qquad & \text{by definition} \\
&= 1 + (\alpha L - \alpha L) + \left(\alpha^2 L^2 - \alpha^2 L^2\right) + \cdots \\
&= \left(1 + \alpha L + \alpha^2 L^2 + \cdots\right) - \left(\alpha L + \alpha^2 L^2 + \cdots\right) \\
&= (1-\alpha L)\left(1 + \alpha L + \alpha^2 L^2 + \cdots\right) \\
&= (1-\alpha L)\sum_{i=0}^{\infty} \alpha^i L^i \,.
\end{aligned}$$

For (1.2),

$$\begin{aligned}
(1-\alpha L)^{-1} &= -(\alpha L)^{-1}\left(1-(\alpha L)^{-1}\right)^{-1} \\
&= (-\alpha L)^{-1}\sum_{i=0}^{\infty}\left((\alpha L)^{-1}\right)^i \text{, from (1.1)} \\
&= -\alpha^{-1}L^{-1}\sum_{i=0}^{\infty}\alpha^{-i}L^{-i} \\
&= -\sum_{i=0}^{\infty}\alpha^{-(i+1)}L^{-(i+1)} = -\sum_{i=1}^{\infty}\alpha^{-i}L^{-i} \,.
\end{aligned}$$

EXAMPLE 1.1. In the equation $c_1 y_t + c_0 y_{t-1} = g_t$, let the *function* g_t be represented by some known *sequence* of real numbers x_t. Then, with $b = c_0 / c_1$, we can write $y_t + b y_{t-1} = (c_1)^{-1} x_t = X_t$. In lag notation, $(1+bL)y_t = X_t$. We then have

$$
\begin{aligned}
y_t = (1+bL)^{-1} X_t = \left(1-(-b)L\right)^{-1} X_t \\
&= \sum_{i=0}^{\infty} (-b)^i L^i X_t \\
&= \sum_{i=0}^{\infty} (-b)^i X_{t-i} \\
&= \frac{1}{c_1} \sum_{i=0}^{\infty} \left(-\frac{c_0}{c_1}\right)^i x_{t-i} \,.
\end{aligned}
$$

If $|b|>1$, the series is divergent, so the use of the series expansion (1.1) will not work. In this case, we make use of (1.2), yielding the solution

$$
y_t = -\frac{1}{c_1} \sum_{i=1}^{\infty} \left(-\frac{c_1}{c_0}\right)^i x_{t+i} \,.
$$

The solution for $|b|<1$ gave a geometrically weighted sum of the current and past values of x_t. That is, we got the *backward solution*. The solution for $|b|>1$ gave a geometrically weighted sum of future values of x_t. That is, we got the *forward solution*. •

The General Solution

At this point, some of you may (quite correctly) be a little puzzled. We began with a homogeneous equation $c_1 y_t + c_0 y_{t-1} = 0$ and derived a solution that depended on an initial condition. We then looked at the inhomogeneous equation $c_1 y_t + c_0 y_{t-1} = g_t$ and our solutions made no mention of initial conditions. What is going on? The answer is that we must distinguish between *general* and *particular* solutions of a difference equation. For the homogeneous equation we had a general solution

$$
y_t = A\left(-\frac{c_0}{c_1}\right)^t ,
$$

for some unknown A. A particular solution in this case is one in which we select A from a set of appropriate boundary conditions. It turns out that for the inhomogeneous equations we have been solving only for a particular solution. However, as we will see shortly, the particular solution we had solved for is special because it corresponds to the *equilibrium solution* of a model.

To find the general solution of the inhomogeneous equation we make use of the following result:

THEOREM: *Consider the equation* $c_1 y_t + c_0 y_{t-1} = g_t$. *Let* $\overline{y}_t$ *denote any particular solution to the inhomogeneous equations, and let* $\tilde{y}_t$ *denote the general solution to the homogeneous equation* $c_1 y_t + c_0 y_{t-1} = 0$. *Then, the general solution to the homogeneous equation is* $y_t = \tilde{y}_t + \overline{y}_t$, *where* $\tilde{y}_t$ *contains an unknown constant. A particular solution can be obtained by solving for the unknown constant term by exploiting boundary conditions.*

EXAMPLE 1.2. For the equation $c_1 y_t + c_0 y_{t-1} = x_t$, $|c_0| < |c_1|$, we have already seen

$$\tilde{y}_t = A\left(-\frac{c_0}{c_1}\right)^t \quad \text{and} \quad \overline{y}_t = \frac{1}{c_1}\sum_{i=0}^{\infty}\left(-\frac{c_0}{c_1}\right)^i x_{t-i} .$$

Thus, the general solution is

$$y_t = A\left(-\frac{c_0}{c_1}\right)^t + \frac{1}{c_1}\sum_{i=0}^{\infty}\left(-\frac{c_0}{c_1}\right)^i x_{t-i} .$$

Now, assume there is a set of initial conditions $x_t = 0 \ \forall t < 0$, x_0=1, y_0=1. Then, we can obtain the particular solution by solving for A:

all zero

$$1 = y_0 = A + \frac{1}{c_1} + \cancel{\frac{1}{c_1}\sum_{i=1}^{\infty}\left(-\frac{c_0}{c_1}\right)^i} x_{t-i} ,$$

which gives $A = (1 - c_1)/c_1$, and

$$y_t = \left(\frac{1-c_1}{c_1}\right)\left(-\frac{c_0}{c_1}\right)^t + \frac{1}{c_1}\sum_{i=0}^{\infty}\left(-\frac{c_0}{c_1}\right)^i x_{t-i} . \qquad \bullet$$

Example 1.2 is useful to now clarify what we meant by the statement that a particular solution can correspond to the *equilibrium solution* of a model. Imagine that the process being analyzed has been in process for a very long time, indexed by t_0, but that we are beginning to observe it only at some time $t+t_0$. Then, the general solution is

$$y_{t+t_0} = A\left(-\frac{c_0}{c_1}\right)^{t+t_0} + \frac{1}{c_1}\sum_{i=0}^{\infty}\left(-\frac{c_0}{c_1}\right)^{i} x_{t+t_0-i}\,.$$

But if the process has been underway for a very long time, $t_0 \to \infty$, and we have

$$\lim_{t_0\to\infty} y_{t+t_0} = \lim_{t_0\to\infty} \frac{1}{c_1}\sum_{i=0}^{\infty}\left(-\frac{c_0}{c_1}\right)^{i} x_{t+t_0-i}\,.$$

The general solution of the homogenous equation vanishes. Now re-index time to s where $s=t_0-t$. Then, we have

$$y_s = \frac{1}{c_1}\sum_{i=0}^{\infty}\left(-\frac{c_0}{c_1}\right)^{i} x_{s-i}\,,$$

where $s=0$ corresponds to the first period we observe the process. That is, the particular solution obtained by setting $\tilde{y}_t = 0$ represents the particular solution for a process that has already been underway for a long period of time before we begin to observe it, so the initial conditions that applied when the process began no longer matter. This is what is meant by an equilibrium process.

Just to cement this idea, consider an inhomogeneous difference equation intended to represent, say, the number, n_t, of firms active in an industry. If we are studying an industry from its birth, when the first firm entered, we obtain the general solution and then select from this a particular solution using the initial condition $n_0=1$. In contrast, if we are studying an industry that is already well-established at the time we begin to study it, we want the equilibrium solution, i.e. the particular solution to the inhomogeneous equation obtained by setting the general solution to the homogeneous equation to zero.

EXAMPLE 1.3 (*Partial adjustment in firm output*). Profit-maximizing output for a firm is given by $\hat{y}_t = f(p_t)$, where industry price, p_t, is some arbitrary function of time. The firm faces a loss when it does not set $y_t = \hat{y}_t$, equal to $(a/2)\left(y_t - \hat{y}\right)^2$, so it would like to set

$y_t = \hat{y}_t$, in every period. However, there is a quadratic cost to adjusting output equal to $(b/2)(y_t - y_{t-1})^2$. Find the output chosen by a myopic cost-minimizer.

The myopic cost-minimizer considers only the current period's costs, and does not think about how today's output choice may influence future costs. That is, the firm's objective is

$$\min_{y_t} L = \frac{a}{2}(y_t - \hat{y}_t)^2 + \frac{b}{2}(y_t - y_{t-1})^2 .$$

The first-order condition yields

$$y_t = \left(\frac{a}{a+b}\right)\hat{y}_t + \left(\frac{b}{a+b}\right)y_{t-1} .$$

This is the difference equation to solve. Before doing so, let us rewrite it as

$$y_t - y_{t-1} = \frac{a}{a+b}(\hat{y}_t - y_{t-1}) . \tag{1.3}$$

This is the famous partial adjustment equation. The change in a firm's output is a fraction $a/(a+b) \in (0,1)$ of the gap between the actual and desired output levels. The greater the cost of not hitting the target relative to the cost of adjustment, the greater the speed of adjustment. Now, rewrite the equation again:

$$(a+b)y_t - by_{t-1} = a\hat{y}_t ,$$

and it is easy to see that this is the equation we just solved in the Example 1.2, with $c_1 = (a+b)$, $c_0 = -b$, and $x_t = a\hat{y}_t$. If we are modeling a firm since birth, with entry size y_0, then we take the general solution and use y_0 to pin down the unknown constant:

$$y_t = y_0\left(\frac{b}{a+b}\right)^t + \frac{a}{a+b}\sum_{i=0}^{t}\left(\frac{b}{a+b}\right)^i \hat{y}_{t-i} ,$$

where the upper limit to the summation is set to t because we don't care about optimal output levels before the firm was born.

If, instead, we are modeling the output path of a firm that has been around a long period of time, it might be reasonable to assume that it has been around for an infinite period of time and thus

$$y_t = \frac{a}{a+b}\sum_{i=0}^{\infty}\left(\frac{b}{a+b}\right)^i \hat{y}_{t-i} \ . \qquad \bullet$$

A Digression: Myopia Versus Farsightedness

While sometimes it is justifiable to assume myopia on the part of firms, the most common reason the assumption is imposed in practice is simplicity. It is the job of the modeler to decide whether myopia can be justified, or whether its analytical convenience outweighs the insights that might be obtained by dropping the assumption.

To illustrate the additional insight and complexity that result from assuming farsightedness we write the objective function as

$$\min_{y_t} L = \sum_{i=0}^{\infty}\beta^i \left\{\frac{a}{2}\left(y_{t+i} - \hat{y}_{t+i}\right)^2 + \frac{b}{2}\left(y_{t+i} - y_{t+i-1}\right)^2\right\},$$

where $\beta<1$ is the discount factor. The firm chooses current output to minimize the discounted sum of the current and all future losses. The first-order condition is

$$a\left(y_t - \hat{y}_t\right) + b\left(y_t - y_{t-1}\right) - \beta b\left(y_{t+1} - y_t\right) = 0 \ ,$$

which can be written in classic partial adjustment form as

$$y_t - y_{t-1} = \left(\frac{a}{a+b}\right)\left(\hat{y}_t - y_{t-1}\right) + \beta\left(\frac{b}{a+b}\right)\left(y_{t+1} - y_t\right). \tag{1.4}$$

Compare (1.4) with the myopic case (1.3). The farsighted firm has an additional term affecting the change in current output. As in the myopic case, the change in a firm's output is a linear function of the fraction $a/(a+b)$ of the gap between the actual and desired output levels. However, current output is further revised by a fraction $\beta b/(a+b) \in (0,1)$ of the difference between this and next period's output. The new term arises because the farsighted firm understands that this period's output choice will affect next period's adjustment costs. The higher the adjustment costs relative to the loss from the output gap, the more weight is applied to this additional term. The lower the discount factor, the less the firm cares about the future consequences of its current actions. Note that (1.4) includes terms involving t–1, t, and t+1, so the partial adjustment model for the farsighted firm generates a *second-order* difference equation. We will see how to solve this type of equation later.

Mapping Bounded Sequences to Bounded Sequences

Consider the equation

$$y_t = \lambda y_{t-1} + bx_t + a, \quad |\lambda| > 1\,.$$

Because $|\lambda| > 1$, we want to use the forward solution, which is

$$y_t = c\lambda^t + \frac{a}{1-\lambda} - b\sum_{i=1}^{\infty} \lambda^{-i} x_{t+i}\ ,$$

where c is a constant yet to be determined. How do we determine this constant if we have no explicit boundary condition? One way is to make the following assumption:

> *The only admissible solutions for any bounded sequence* $\{x_t\}_{t=0}^{\infty}$ *consist of bounded sequences* $\{y_t\}_{t=0}^{\infty}$.

That is, if the exogenous variable is bounded, then we assume the endogenous variable is also. In this example, the only way to generate a bounded sequence for y is to set c=0. Of course, it is the economic model that must justify the assumption. For example, if x is price and y is industry output, it may well be reasonable to restrict attention to finite output levels whenever price is finite.[2]

2. Rational Expectations and Uncertainty

In this section we extend our analysis of linear difference equations to admit (i) uncertainty and (ii) the assumption that expectations are formed rationally, in the traditional sense that our modeling of expectations sets agent's expectations equal to the conditional mathematical expectation implied by the model.

---

[2] Unfortunately, the assumption may outlaw some solutions without yielding a unique solution. Indeed, we may be left with an infinite number of admissible solutions. This depends on the model.

Expectations Rules

The following properties of random variables may be useful. Let

$$z = a + bx, \quad w = c + dy\ ,$$

where x, y, w, and z are random variables, and a, b, c, and d are constants.

- EXPECTED VALUE:

$$E(x+y) = E(x) + E(y) \qquad\qquad E(z) = a + bE(x)\,.$$

- VARIANCES:

$$\operatorname{var}(x) = E\Big[\big(x - E(x)\big)^2\Big] = E\left(x^2\right) - \big[E(x)\big]^2\ ;$$

$$\operatorname{cov}(x,y) = E\Big[\big(x - E(x)\big)\big(y - E(y)\big)\Big] = E(xy) - E(x)E(y)\ ;$$

$$\operatorname{var}(z) = b^2 \operatorname{var}(x) \qquad\qquad \operatorname{var}(x+y) = \operatorname{var}(x) + \operatorname{var}(y) + 2\operatorname{cov}(x,y)$$

$$\operatorname{cov}(z,w) = bd \operatorname{cov}(x,y) \qquad\qquad \operatorname{var}(x-y) = \operatorname{var}(x) + \operatorname{var}(y) - 2\operatorname{cov}(x,y)$$

- CONDITIONAL EXPECTATIONS:

Let I_t denote information available at time t, and let $\omega \in I_t$ be a subset of that information. Then the expectation of x conditional on the information set I_t is denoted by $E\left[x \mid I_t\right]$, or more informally by $E_t\,(x)$. The *law of iterated expectations* is

$$E\Big[E\left[x \mid I_t\right] \mid \omega_t\Big] = E\left[x \mid \omega_t\right].$$

Heuristically, the law of iterated expectations means that if I am asked how I would revise my expectations if I were given more information, my answer would be that they are equally likely to go up or down and on average my expectation is that the revision would be zero. In particular, the law of iterated expectations means

$$E\Big[E\left[x \mid I_{t+1}\right] \mid I_t\Big] = E\left[x \mid I_t\right],$$

and we will use this frequently in solving linear rational expectations models. Note that adaptive expectations, studied earlier in this course, do not satisfy the law of iterated expectations.

• AUTOREGRESSIVE PROCESSES:

Certain autoregressive stochastic processes appear commonly in linear rational expectations models. We will see examples of AR(1) [first-order autoregressive] processes taking the form

$$x_t = \alpha x_{t-1} + \varepsilon_t \,,$$

where ε_t is a *white noise* random variable, uncorrelated across time periods with zero mean and constant variance σ_ε^2. We can derive the following relationships,

$$\operatorname{cov}\left(x_t, x_{t-1}\right) = \alpha^i \sigma_x^2 \qquad \sigma_x^2 = (1/(1-\alpha^2))\sigma_\varepsilon^2 \qquad E\left[x_{t+j} \mid x_t\right] = \alpha^j x_t \,,$$

noting that the variance of x is finite only for $|\alpha|<1$. When $\alpha=1$, the AR(1) process is known as a *random walk*. In the case of a random walk, we have $E\left[x_{t+j} \mid x_t\right] = x_t \;\; \forall j > 0$.

Solving Linear Rational Expectations Models

There are several techniques for solving linear rational expectations models. We will use three techniques to solve the following simple model:

$$y_t = aE\left[y_{t+1} \mid I_t\right] + cx_t \,, \tag{2.1}$$

where x_t is an exogenous, stochastic process. In modeling how expectations are formed, we make the following assumptions:

• Agents know the model, and the values of a and c.

• All agents have the same information set, so we can talk about "the" mathematical expectation based on "the" information set.

• When forming expectations at time t, the agent has observed x_t. That is, $x_t \in I_t$.

In this section, we restrict attention to the following conditions:

• $a \in (0,1)$

• y_t is bounded.

These conditions provide us with a unique solution to (2.1). We will drop them later to see what happens.

Method 1: Repeated Substitution. Rewrite the problem for $t+1$:

$$y_{t+1} = aE\left[y_{t+2} \mid I_{t+1}\right] + cx_{t+1} ,$$

and take expectations conditional on I_t:

$$E\left[y_{t+1} \mid I_t\right] = aE\left[y_{t+2} \mid I_t\right] + cE\left[x_{t+1} \mid I_t\right],$$

where, in the first term on the right hand side, we have applied the law of iterated expectations. Now substitute this expression for $E\left[y_{t+1} \mid I_t\right]$ into (2.1):

$$y_t = a^2E\left[y_{t+2} \mid I_t\right] + acE\left[x_{t+1} \mid I_t\right] + cx_t .$$

Repeat this substitution up to time $t+T$:

$$y_t = c\sum_{i=0}^{T} a^i E\left[x_{t+i} \mid I_t\right] + a^{T+1}E\left[y_{t+T+1} \mid I_t\right].$$

Now, if y_t is bounded, then as $|a|<1$ we have

$$\lim_{T\to\infty} a^{T+1}E\left[y_{t+T+1} \mid I_t\right] = 0 , \tag{2.2}$$

and so,

$$y_t = c\sum_{i=0}^{\infty} a^i E\left[x_{t+i} \mid I_t\right]. \tag{2.3}$$

Note that y_t can only be bounded if $E\left[x_{t+i} \mid I_t\right]$ does not grow too fast. Specifically, we requires that

$$\frac{E\left[x_{t+i+1} \mid I_t\right]}{E\left[x_{t+i} \mid I_t\right]} < \frac{1}{a} ,$$

or equivalently that the growth rate of $E\left[x_{t+i} \mid I_t\right]$ is less than $(1/a)$–1. Equation (2.3) is the unique solution to the problem under assumption (2.3). If we are willing to make some assumptions about the expected time path for x_t, we can solve more explicitly for y_t, as the following examples show.

EXAMPLE 2.1. Assume $x=x_0$ for $t\leq T$, and $x=x_1$ for $t>T$, and this path is known to the agent we are modeling. Then, we have

$$y_t = c\sum_{i=0}^{\infty} a^i E\left[x_{t+i} \mid I_t\right]$$

$$= c\sum_{i=0}^{T} a^i x_0 + c\sum_{i=T+1}^{\infty} a^i x_1$$

$$= cx_0\left(\frac{1-a^{T+1}}{1-a}\right) + cx_1\left(\frac{a^{T+1}}{1-a}\right) \qquad \bullet$$

EXAMPLE 2.2. Assume $x_{t+1} = x_t + \varepsilon_{t+1}$, where ε is independently and identically distributed with mean zero. Then, $E\left[x_{t+i} \mid I_t\right] = x_t$ for all i, so

$$y_t = c\sum_{i=0}^{\infty} a^i x_t = \frac{cx_t}{1-a}. \qquad \bullet$$

Method 2: Undetermined Coefficients. As in the deterministic case, we guess a functional form for the solution and then verify it. Let us guess a form for the solution that we already know is correct. Guess

$$y_t = \sum_{i=0}^{\infty} \lambda_i E\left[x_{t+i} \mid I_t\right], \tag{2.3}$$

where λ_i, i=1, 2, 3, . . . , are coefficients to be determined. If this guess is correct, then imposing rational expectations gives us

$$E\left[y_{t+1} \mid I_t\right] = \sum_{i=0}^{\infty} \lambda_i E\left[x_{t+i+1} \mid I_t\right]. \tag{2.4}$$

We now substitute our guesses, (2.3) and (2.4) into the original equation (2.1), to obtain

$$\sum_{i=0}^{\infty} \lambda_i E\left[x_{t+i} \mid I_t\right] = a\sum_{i=0}^{\infty} \lambda_i E\left[x_{t+i+1} \mid I_t\right] + cx_t.$$

This equation should hold for *any* realizations of the sequence $\left\{x_{t+i}\right\}_{i=1}^{\infty}$, and the only way this can happen is if for every i, the coefficient on x_{t+i} on the left hand side of the equation is identical to the coefficient on x_{t+i} on the right hand side. Matching up the coefficients, we get

$$\lambda_0 = c, \text{ and } \lambda_{i+1} = a\lambda_i = a^i c,$$

and this again yields (2.3).

EXERCISE 2.1 (A supply and demand model). *Solve the following problem using the method of undetermined coefficients.*

$$y_t = \alpha_0 - \alpha_1 p_t + e_t\ ,$$
$$y_t = \beta_0 + \beta_1 E\left[p_t \mid I_{t-1}\right],$$
$$e_t = \rho e_{t-1} + \varepsilon_t\ , \qquad E\left[\varepsilon_t = 0\right].$$

(***S****olve for price first, assuming a solution of the form* $p_t = \pi_0 + \pi_1\varepsilon_t + \pi_2 e_{t-1}$).

Method 3: Sargent's Factorization Method. We have seen the use of lag and forward operators already. Their use in solving stochastic rational expectations models was pioneered by Sargent (1979). With the introduction of expectations operators, it is important to note that the lag and forward operators work on the time-subscript of the variable and *not* on the time subscript of the information set. That is,

$$LE\left[p_{t+i} \mid I_{t-1}\right] \equiv E\left[p_{t+i-1} \mid I_{t-1}\right],$$
$$LE\left[p_{t+1} \mid I_t\right] \equiv E\left[p_t \mid I_t\right] \equiv p_t\,.$$

Consider our simple problem, (2.1). Sargent's factorization method first involves taking expectations on both sides of the equations conditional on the oldest information set that appears anywhere in the equation. In this simple problem, there is only one information set, I_t, so we take expectations over the entire equation based on I_t:

$$E\left[y_t \mid I_t\right] = aE\left[y_{t+1} \mid I_t\right] + cE\left[x_t \mid I_t\right]. \tag{2.5}$$ [3]

The second step in Sargent's method is to write (2.5) in terms of the operators:

$$E\left[y_t \mid I_t\right] = aFE\left[y_t \mid I_t\right] + cE\left[x_t \mid I_t\right],$$

which implies

---

[3] Note that in this particular example, we have $y_t \in I_t$ and $x_t \in I_t$, so $E\left[y_t \mid I_t\right] = y_t$, $E\left[x_t \mid I_t\right] = x_t$, and $y_t = aE\left[y_{t+1} \mid I_t\right] + cx_t$. That is, the first step in Sargent's method does nothing in this example because there was no lagged information set, such as I_{t-1}. We will see examples later when this step does something substantial to the equation.

$$(1-aF)\,E\big[y_t \mid I_t\big] = cE\big[x_t \mid I_t\big]\,,$$

or

$$\begin{aligned} E\big[y_t \mid I_t\big] &= c\,(1-aF)^{-1}\,E\big[x_t \mid I_t\big] \\ &= c\sum_{i=0}^{\infty} a^i F^i E\big[x_t \mid I_t\big]\,, \end{aligned}$$

which is the same solution as before.

This problem was simple, and all the methods were easy to apply. In general, is there any preference? For the simplest problems, repeated substitution is the most straightforward. However, it quickly becomes unwieldy as the problem increases in complexity. Undetermined coefficients eventually also become awkward because the initial guess may not include the correct solution. Sargent's method is the most powerful, particularly for problems with multiple solutions. However, it is often found to be the most conceptually challenging.

Multiple Solutions

Difference equations with order greater than one usually have multiple solutions. In the case where there are just a few solutions, then selecting among them can usually be accomplished by choosing the solution that is stable (i.e. the one in which there is an adjustment process that does not lead to exploding values for the endogenous variables(s). We will see how this is done by means of two examples.

EXAMPLE 2.3. Consider the following supply and demand model

$$y_t = -\alpha_0 p_t + \alpha_1\Big(E\big[p_{t+1} \mid I_t\big] - p_t\Big) + e_t\,, \tag{2.6}$$

$$y_t = \beta_0 E\big[p_t \mid I_{t-1}\big] + \beta_1\Big(E\big[p_t \mid I_{t-1}\big] - p_{t-1}\Big) \tag{2.7}$$

where e_t is white noise, and the four constant coefficients are all positive. Equation (2.6) is the demand curve, which contains a speculative component: demand is high in period t if agents expect a large price increase by period t+1. The supply curve, (2.7), contains a lag in production and a speculative component. The production lag is reflected in the fact that supply decision are made on the basis of I_{t-1}. The speculative component states that supply is high in period t if in period t–1 agents had expected the price to be much higher than it was in period t–1. This component comes, of course, from the previous period's speculative demand.

Combining (2.6) and (2.7) gives us a single equation in prices:

$$\beta_0 E\left[p_t \mid I_{t-1}\right] + \beta_1\left(E\left[p_t \mid I_{t-1}\right] - p_{t-1}\right) = -\alpha_0 p_t + \alpha_1\left(E\left[p_{t+1} \mid I_t\right] - p_t\right) + e_t\,. \tag{2.8}$$

We will solve this by the method of undetermined coefficients. It turns out that the solution has the form

$$p_t = \pi_1 e_t + \pi_2 p_{t-1}\,, \tag{2.9}$$

although some comments about guessing the functional form are in order. First, note that (2.8) has no intercept, so none is included in the guessed functional form. Second, our most general guess should take the form

$$\begin{aligned} p_t &= \sum_{i=-\infty}^{+\infty} \pi_{1i} E\left[e_{t-i} \mid I_t\right] + \sum_{i=1}^{\infty} \pi_{2i} p_{t-i}\,. \\ &= \sum_{i=0}^{\infty} \pi_{1i} e_{t-1} + \sum_{i=1}^{\infty} \pi_{2i} p_{t-i}\,. \end{aligned}$$

The tactic here is to include all past values of the endogenous variable, and all values of the exogenous variables that appear in (2.8). Because e_t is white noise, $E\left[e_{t+i} \mid I_t\right] = 0$ for all i>0. It also turns out in this case that $\pi_{1i} = 0$ for all i>0, and $\pi_{2i} = 0$ for all i>1, so we will save time and assume (2.9).

Imposing rational expectations on the guessed functional form gives

$$E\left[p_t \mid I_{t-1}\right] = \pi_2 p_{t-1}\,, \tag{2.10}$$

$$E\left[p_{t+1} \mid I_t\right] = \pi_2 p_t = \pi_1 \pi_2 e_t + \pi_2^2 p_{t-1}\,, \tag{2.11}$$

and substituting (2.9)–(2.11) into (2.8) yields

$$\begin{aligned} &\left(-\beta_0 \pi_2 - \beta_1 \pi_2 - \alpha_0 \pi_2 - \alpha_1 \pi_2^2 - \alpha_1 \pi_2 + \beta_1\right) p_{t-1} \\ &\qquad + \left(-\alpha_0 \pi_1 + \alpha_1 \pi_1 \pi_2 - \alpha_1 \pi_1 + 1\right) e_t = 0\,. \end{aligned} \tag{2.12}$$

This equation must hold for any values of e_t and p_{t-1}. Thus, the terms inside parentheses must be zero. The first term then gives a quadratic solution for the unknown coefficient π_2:

$$\pi_2 = \frac{\alpha_0 + \beta_0 + \alpha_1 + \beta_1}{2\alpha_1} \pm \frac{1}{2\alpha_1}\sqrt{\left(\alpha_0 + \beta_0 + \alpha_1 + \beta_1\right) - 4\alpha_1\beta_1} \,. \tag{2.13}$$

We have two possible solutions for π_2. Which one is economically interesting? Generally, we will find that only one corresponds to a stable solution. Take a look again at the posited solution, (2.9). A shocks to p will dampen down over time only if $\left|\pi_2\right| < 1$. Thus, we look for a solution with $\left|\pi_2\right| < 1$, and it turns out that only one solution can satisfy this stability requirement. To see this, we apply a neat little trick for quadratic solutions. Let π_2^- and π_2^+ denote the two solutions to (2.13) and, without loss of generality, assume $\pi_2^- \le \pi_2^+$. Then, we can easily verify that

$$\pi_2^- \pi_2^+ = \frac{\beta_1}{\alpha_1} > 0 \tag{2.14}$$

and

$$\pi_2^- + \pi_2^+ = \frac{\alpha_0 + \beta_0 + \alpha_1 + \beta_1}{2\alpha_1} > 0 \,. \tag{2.15}$$

Equation (2.14) shows that both solutions have the same sign. This fact, when used with (2.15) shows that both roots are positive, so we have

$$0 < \pi_2^- \le \pi_2^+ \,.$$

So far, we still don't know whether we have zero, one, or two stable solutions. To get any further, we need to do some *economic* thinking. For this problem, note that α_1 should be equal to β_1. This is because α_1 is the coefficient of speculative demand and measures how prices affect the quantity that is bought today in order to sell tomorrow; β_1 is the coefficient on the inventory carryover of speculators' supply and measures how prices affected the quantity that was bought yesterday in order to sell today. These two quantities should logically be equally sensitive to price changes. Imposing equality of coefficients, (2.15) tells us that

$$\pi_2^- = \frac{1}{\pi_2^+}$$

which tells us that $0 < \pi_2^- \leq 1$ and $1 \leq \pi_2^+$. If $\left(\alpha_0 + \beta_0 + \alpha_1 + \beta_1\right) - 4\alpha_1\beta_1$, then there is a single solution with $\pi_2^- = \pi_2^+ = 1$. More generally, there will be a single stable solution,

$$\pi_2^- = \frac{\alpha_0 + \beta_0 + \alpha_1 + \beta_1}{2\alpha_1} - \frac{1}{2\alpha_1}\sqrt{\left(\alpha_0 + \beta_0 + \alpha_1 + \beta_1\right) - 4\alpha_1\beta_1}\,. \tag{2.16}$$

Returning to (2.12), we can see that

$$\pi_1 = \frac{1}{\alpha_0 + \alpha_1\left(1 - \pi_2\right)},$$

so that

$$p_t = \frac{1}{\alpha_0 + \alpha_1\left(1 - \pi_2\right)} e_t + \pi_2 p_{t-1},$$

where π_2 is as defined in (2.16). Now we have the solution for p_t, obtaining the solution for y_t is straightforward and this is left as an exercise.

Although the demand disturbances are pure white noise, the presence of speculation introduces positive serial correlation in prices. The causal mechanism is intuitive. If e_{t-1} is high, then p_{t-1} will also be high. Speculators will cut back their purchases, and this will reduce speculative supply in period t. Hence, p_t will also be higher than normal. •

EXAMPLE 2.4. We will solve the equation,

$$p_t = a_0 E\left[p_{t+1} \mid I_t\right] + a_1 p_{t-1} + a_2 E\left[p_t \mid I_{t-1}\right] + a_3 y_t + e_t\,,$$

where all coefficient are positive, by Sargent's factorization method. The first step is to take expectations of the entire equation based on the oldest information set, I_{t-1}:

$$E\left[p_t \mid I_{t-1}\right] = a_0 E\left[p_{t+1} \mid I_{t-1}\right] + a_1 p_{t-1} + a_2 E\left[p_t \mid I_{t-1}\right] + E\left[x_t \mid I_{t-1}\right], \tag{2.17}$$

where, to ease notation, we have written $x_t = a_3 y_t + e_t$. In writing (2.17), we again applied the law of iterated expectations and assumed that $p_{t-1} \in I_{t-1}$. For convenience, rearrange this equation slightly:

$$\left(1 - a_2\right) E\left[p_t \mid I_{t-1}\right] = a_0 E\left[p_{t+1} \mid I_{t-1}\right] + a_1 p_{t-1} + E\left[x_t \mid I_{t-1}\right],$$

which, in terms of lag and forward operators can be written as

$$(1-a_2)E[p_t \mid I_{t-1}] = a_0 FE[p_t \mid I_{t-1}] + a_1 Lp_t + E[x_t \mid I_{t-1}].$$

We can collect terms involving p_t, noting that $Lp_t = LE[p_t \mid I_{t-1}]$, to obtain

$$(-a_0 F + (1-a_2) - a_1 L)E[p_t \mid I_{t-1}] = E[x_t \mid I_{t-1}],$$

or, upon dividing through by $-a_0$,

$$\left(F^2 + \left(\frac{1-a_2}{a_0}\right)F + \frac{a_1}{a_0}\right)LE[p_t \mid I_{t-1}] = -\frac{1}{a_0}E[x_t \mid I_{t-1}]. \tag{2.18}$$

or,

$$(F-\lambda_1)(F-\lambda_2)LE[p_t \mid I_{t-1}] = -\frac{1}{a_0}E[x_t \mid I_{t-1}], \tag{2.19}$$

where

$$\lambda_1 + \lambda_2 = \frac{1-a_2}{a_0} \text{ and } \lambda_1\lambda_2 = \frac{a_1}{a_0}. \tag{2.20}$$

To see this last step, expand $(F-\lambda_1)(F-\lambda_2)$ and compare terms in (2.18) and (2.20). This factoring in (2.19) explains, at last, the meaning of the term "Sargent's factorization method". From (2.20), we can solve for

$$\{\lambda_1, \lambda_2\} = \frac{1-a_2}{a_0} \pm \frac{1}{2}\sqrt{\frac{(1-a_2)^2}{a_0^2} - \frac{4a_1}{a_0}}.$$

The values of λ_1 and λ_2 depend on the values of the parameters of the model. We will assume that $0 < \lambda_1 < 1 < \lambda_2$, which turns to be the assumption necessary for a saddle-path stable solution in the sense described in the chapter on differential equations. (We are about to use the series expansions (1.1) and (1.2); normally, we would check our assumptions about the admissible range of values to ensure that the series expansions converge. In practice, it is often only at this stage that the modeler first discovers that some restrictions on the parameter values must be imposed).

From (2.19), we can write

$$(1-\lambda_1 L)(F-\lambda_2)E[p_t \mid I_{t-1}] = -\frac{1}{a_0}E[x_t \mid I_{t-1}]$$

and

$$\left(1-\lambda_1 L\right) E\left[p_t \mid I_{t-1}\right] = -\frac{1}{a_0}\left(F-\lambda_2\right)^{-1} E\left[x_t \mid I_{t-1}\right]$$
$$= \frac{1}{a_0\lambda_2}\left(1-\lambda_2^{-1}F\right)^{-1} E\left[x_t \mid I_{t-1}\right]$$

and since $0 < 1/\lambda_2 < 1$, we can expand the right hand side into the convergent series,

$$\left(1-\lambda_2^{-1}F\right)^{-1} = \sum_{i=0}^{\infty} \lambda_2^{-i} F^i ,$$

so that

$$E\left[p_t \mid I_{t-1}\right] = \lambda_1 L E\left[p_t \mid I_{t-1}\right] + \left(\frac{1}{a_0\lambda_2}\right)\sum_{i=0}^{\infty} \lambda_2^{-i} F^i E\left[x_t \mid I_{t-1}\right]$$
$$= \lambda_1 p_{t-1} + \left(\frac{1}{a_0\lambda_2}\right)\sum_{i=0}^{\infty} \lambda_2^{-i} E\left[x_{t+i} \mid I_{t-1}\right]. \tag{2.21}$$

The last step is to derive the solution for p_t itself. Updating (2.21) by one time period, we obtain

$$E\left[p_{t+1} \mid I_t\right] = \lambda_1 p_t + \left(\frac{1}{a_0\lambda_2}\right)\sum_{i=0}^{\infty} \lambda_2^{-i} E\left[x_{t+i+1} \mid I_t\right], \tag{2.22}$$

and substituting (2.21) and (2.22) into the original problem yields

$$p_t = a_0\lambda_1 p_t + \left(\frac{1}{\lambda_2}\right)\sum_{i=0}^{\infty} \lambda_2^{-i} E\left[x_{t+i+1} \mid I_t\right] + a_1 p_{t-1}$$
$$+ a_2\lambda_1 p_{t-1} + \left(\frac{a_2}{a_0\lambda_2}\right)\sum_{i=0}^{\infty} \lambda_2^{-i} E\left[x_{t+i} \mid I_{t-1}\right] + x_t .$$

Rearranging, using (2.20) to simplify, gives

$$p_t = \lambda_1 p_{t-1} + \left(\frac{1}{1-a_0\lambda_1}\right)\sum_{i=0}^{\infty} \lambda_2^{-i} E\left[x_{t+i} \mid I_t\right] + \left(\frac{a_2}{a_0\left(1-a_0\lambda_1\right)}\right)\sum_{i=0}^{\infty} \lambda_2^{-i-1} E\left[x_{t+i} \mid I_{t-1}\right].$$

The first summation term has absorbed the x_t term.

Note that if we are willing to place restrictions on the exogenous stochastic processes, we may be able to write a more compact solution. For example, assume that e_t is white noise and y_t is a random walk. Then $E\left[x_{t+i} \mid I_t\right] = y_t$ for all i>0, and we get

$$p_t = \lambda_1 p_{t-1} + \left(\frac{b_1}{1 - a_0\lambda_1}\right) y_t + \left(\frac{a_2 b_2}{a_0\left(1 - a_0\lambda_1\right)}\right) y_{t-1}\,,$$

where λ_1, $b_1 = \sum_{i=0}^{\infty} \lambda_2^{-1}$, and $b_2 = \sum_{i=0}^{\infty} \lambda_2^{-(i+1)}$ are constants defined by (2.20). •

Yet More on Multiple Solutions

Rational expectations models can admit many, many, solutions. This is not unique to the difference equations modeled here. Consider the following two-player game repeated over T periods.

		PLAYER 1	
		Action A	Action B
PLAYER 1	Action A	1,1	0,0
	Action B	0,0	2,2

There are two equilibrium strategies, {A,A} and {B,B}. For a T–period game, the solution is found by backward induction. In period T, the equilibrium is either {A,A} or {B,B}, and the same is true in periods T–1, T–2, and so on. This game thus has 2^T equilibrium paths. Game theory has developed numerous concepts to try and select among multiple equilibria – so-called *equilibrium refinements*. In rational expectations models with multiple equilibria, the same problem of selection exists.

To think about the possibilities for equilibrium refinement in stochastic difference equations, consider the following equation for the price of a product:

$$E\left(p_{t+1} \mid I_{t-1}\right) = \alpha + \beta p_t + E\left[p_t \mid I_{t-1}\right] + u_t\,,$$

where u is white noise with constant variance σ_u^2 which links the current price to two expected prices. This equation comes from a well-known rational expectations model due to Taylor (1979). The solution to this model is given by

$$p_t = -\frac{\alpha}{\beta} - \left(\frac{1}{\beta}\right) u_t + \pi \sum_{i=0}^{\infty} (1-\beta)^i u_{t-i-1}\,, \tag{2.23}$$

where π is not only unknown, it is also undetermined. That is, any value of π is a solution. This means that for any value of π, there is a distinct behavior of prices. Moreover, the expectations that agents form will always be consistent with this distinct behavior.

Note that (2.23) implies

$$\operatorname{var}(p_t) = \left[\left(\frac{1}{\beta}\right)^2 + \pi^2 \sum_{i=0}^{\infty} (1-\beta)^{2i}\right] \sigma_u^2\,.$$

One way of selecting among the solutions is to focus on admissible properties of the variance of price. For example, one could require that the variance be finite. This is sometimes, but not always, a reasonable requirement, and it is not always sufficient. If $\beta{>}0$, or if $\beta{<}{-2}$, then $\operatorname{var}(p_t)$ finite requires $\pi{=}0$, and we get a unique solution. But if $-2{<}\beta{<}0$, then $|1{+}\beta|{<}1$ and $\operatorname{var}(p_t)$ is finite for any finite π. In this case, an infinite number of finite-variance rational expectations solutions exist. Taylor (1979) resolved the equilibrium selection problem by strengthening the requirement that $\operatorname{var}(p_t)$ be finite to the condition that it be minimized. This requires that $\pi{=}0$. But what is the justification for this? Moreover, imagine a model with both output and price. Which variance should be minimized?

McCallum (1983) suggested a criterion based on minimizing the set of state variables that must be employed to make a rational expectations forecast. This is arbitrary, but it has intuitive appeal because it recognizes that collecting information is costly and agents may seek to economize on its use. A one-period forecast in this model is, from (2.23),

$$\begin{aligned} E\left[p_{t+1} \mid I_t\right] &= -\frac{\alpha}{\beta} - \left(\frac{1}{\beta}\right) E\left[u_{t+1} \mid I_t\right] + \pi \sum_{i=0}^{\infty} (1-\beta)^i u_{t-i} \\ &= -\frac{\alpha}{\beta} + \pi \sum_{i=0}^{\infty} (1-\beta)^i u_{t-i}\,. \end{aligned}$$

To make this forecast in the general case, the agent must keep track of all the past shocks. The minimum state space model is found where $\pi{=}0$, in which case the forecast is $-\alpha/\beta$. Two objections to McCallum's minimum state space representation are that (i) sometimes the procedure only yields complex-valued solutions, even though real-valued solutions exist, and (ii) sometimes the selected solution does not make economic sense.

EXERCISE 2.2 (Market equilibrium with price-setting firms). *Let demand be given by*

$$y_t = -p_t + aE\left[p_{t+1} \mid I_t\right] + u_t\,,$$

$$u_t = \rho u_{t-1} + \varepsilon_t\,,\ \ 0<a<1,\, 0<\rho<0.$$

and let aggregate supply be given by

$$p_t = cE\left[y_t \mid I_{t-1}\right].$$

The demand equation says that demand will be high if (i) *it is subject to a positive shock, and* (ii) *prices are expected to rise next period. The supply equation says that firms will raise prices if they expect output to be high. The coefficient c can be interpreted as a measure of price flexibility. Let* $\overline{y}_t = E\left[y_t \mid I_{t-1}\right]$ *be the one-period conditional expectation of output, and let* $\tilde{y}_t = y_t - \overline{y}_t$ *be the unanticipated component of output. Show that an increase in flexibility reduces the variance of* $\overline{y}_t$ *but increases the variance of* $\tilde{y}_t$.

EXERCISE 2.3 (Market equilibrium with partial adjustment). C*onsider the following supply and demand system*:

$$D_t = \alpha_0 - \alpha_1 p_t + \varepsilon_t\,,$$

$$S_t = S_{t-1} + \gamma\left(S_t^* - S_{t-1}\right),$$

$$S_t^* = \beta_0 + \beta_1 E\left[p_t \mid I_{t-1}\right],$$

where S_t *is quantity supplied and* $S_t{}^*$ *is the quantity that firms would like to supply.* (a) *Derive the basic difference equation for price.* (b) *Assume that expectations are adaptive, with the form*

$$E\left[p_t \mid I_{t-1}\right] = E\left[p_{t-1} \mid I_{t-2}\right] + \phi\left(p_{t-1} - E\left[p_{t-1} \mid I_{t-2}\right]\right).$$

Solve for price as a function of lagged supply, lagged prices, and the disturbance term. (c) *Assume now that expectations are rational. Solve again for price as a* *function of lagged supply, lagged prices, and the disturbance term.* (d) *Compare your answers in* (b) *and* (c).

EXERCISE 2.4 Consider the following model

$$\begin{aligned} D_t &= -\beta p_t && \text{(demand)} \\ S_t &= \gamma E\left[p_t \mid I_{t-1}\right] + \varepsilon_t && \text{(supply)} \\ I_t &= \alpha\left(E\left[p_{t+1} \mid I_t\right] - p_t\right) && \text{(inventory demand)} \\ S_t &= D_t + \left(I_t - I_{t-1}\right) && \text{(market clearing)} \end{aligned}$$

Suppose there is perfect foresight. This implies that $E\left[p_{t+j} \mid I_t\right] = p_{t+j}$ *for all j. In other words, agents in the model can predicts the time path of the exogenous shocks,* ε_t *with perfect accuracy.* (a) *Write the basic difference equation for price.* (b) *Use the* ***method of factorization*** *to solve for* p_t *as a function of all past and future values of* εt*. Hint: Depending on how you decide to present the solution, you may be able to make use of the following relationship, which holds for* $\lambda \neq 1$,

$$\frac{1}{(1-\lambda L)(1-\lambda^{-1}L)} = \left(\frac{1}{\lambda - \lambda^{-1}}\right)\left(\frac{\lambda}{1-\lambda L} - \frac{\lambda^{-1}}{1-\lambda^{-1}L}\right).$$

EXERCISE 2.5 (Partial adjustment, again). *Recall the partial adjustment equations under rational expectations from section 1:*

$$y_t - y_{t-1} = \left(\frac{a}{a+b}\right)\left(\hat{y}_t - y_{t-1}\right) + \beta\left(\frac{b}{a+b}\right)\left(y_{t+1} - y_t\right). \tag{1.4}$$

Solve it.

3. Nonlinear Difference Equations

Nonlinear equations do not lend themselves to the direct analytical techniques we have looked at for linear equations. In this section, we will begin analyzing equations of the form

$$x_t = f\left(x_{t-1}\right), \tag{3.1}$$

where x_t may be scalar or a vector, so that in the latter case it represents a system of nonlinear equations. Equation (3.1) may give rise to explosive growth (positive or negative), so that $\lim_{t\to\infty}\left|x_t\right| = \infty$, or it may arrive at a finite-valued steady state, in which $x_t = x_{t-1} = x^*$. If such a steady state exists, then

$$x^* = f\left(x^*\right) \tag{3.2}$$

is a **fixed point** of (3.1). When $f(x)$ leads to explosive growth, there are no solutions to (3.2). There may also be multiple solutions.

In analyzing nonlinear difference equations, we will usually be concerned with the following questions:

- Are there fixed points of the equation? If so, how many?
- Which of the fixed points are stable? That is, if we begin at $x_0 \neq x^*$, is the system attracted to x^*?

Naturally, there is an enormous amount of material one could cover. So how are we going to make our selection? Note that the solution of (3.1) in terms of an initial condition is, in contrast to the case for differential equations, a conceptually trivial matter. Iterate on (3.1) repeatedly to get

$$x_t = f\left(x_{t-1}\right) = f\left(f\left(x_{t-2}\right)\right) = \quad \cdot \quad \cdot \quad \cdot \quad = f^{(t)}\left(x_0\right),$$

where $f^{(t)}(\bullet)$ denotes the t^{th} iteration of the function. Often, this is unduly messy, in which case one would want to approach the problem numerically. In rare cases, though, it might be quite easy. For example, if $f(x_t) = x_{t-1}^{\beta}$, then $x_t = \left(x_0\right)^{\beta^t}$. Thus we will restrict our attention in this section to analyses of the existence and stability of fixed points in the limit as $t \to \infty$.

Second, we will make only passing reference to equations that in the limit yield cyclical solutions. This is a matter of time constraint only. A large part of the interest in difference equations is their ability to generate increasingly complex periodic solutions in which x exhibits fluctuations forever. In fact, these cyclical solutions may end up being so complex that they appear to generate entirely unpredictable behavior; some equations in fact actually are so complex that their long run behavior is said to be **chaotic**. But while chaos theory can be fun to study, we are going to put it to one side.

Third, we leave for your own future study the analysis of systems of difference equations. The analysis involves a discrete time analog to the two-dimensional phase diagrams studied for difference equations. While there are important differences between the continuous and discrete time methods, time limitations preclude us from analyzing the latter here. Useful references for these omitted materials are provided at the end of the chapter.

Finally, we note that turning our attention to the study of fixed points will also prove to be particularly useful in preparation for the module on dynamic programming. So this section can be viewed both as an overview of some important aspects of difference equations, and as a technical preparation for the study of discrete time optimization problems.

Steady state properties of scalar equations

Given the one-dimensional function $x_t = f\left(x_{t-1}\right)$, we can we plot x_t as a function of x_{t-1} on a phase diagram. This we have done in Figure 3.1. Using the phase diagrams is straightforward. Given an initial value x_0, we can read the value of x_1 directly from the plotted curve f. To update one period, we move *horizontally* to the 45^0 line, so that x_t on the vertical axis now becomes x_{t-1} on the horizontal axis. To get the value of x for the next period, we read off the value from the curve again. The arrows indicate this evolution of x_t.

In panel (a), there is one fixed point, x^*, and it is stable: any initial value eventually leads us to x^*. In panel (b), the growth rate of x is explosive: there is no fixed point. In panel (c) there are two fixed points, but only the zero fixed point is stable. Finally, in panel (d) there are again two fixed points, but only the larger one is stable.

In Figure 3.1, the function f has been drawn as monotonically increasing in all four panels. While this is a common feature of economic models, it is by no means a universal feature. Figure 3.2 provides some different examples. In panel (a) $f(x)$ is monotonically decreasing, and there is a single stable fixed point. However, note that the fixed point is approached cyclically. In panel (b), again with $f(x)$ monotonically decreasing, the unique fixed point is unstable and x_t exhibits an explosive cycle.

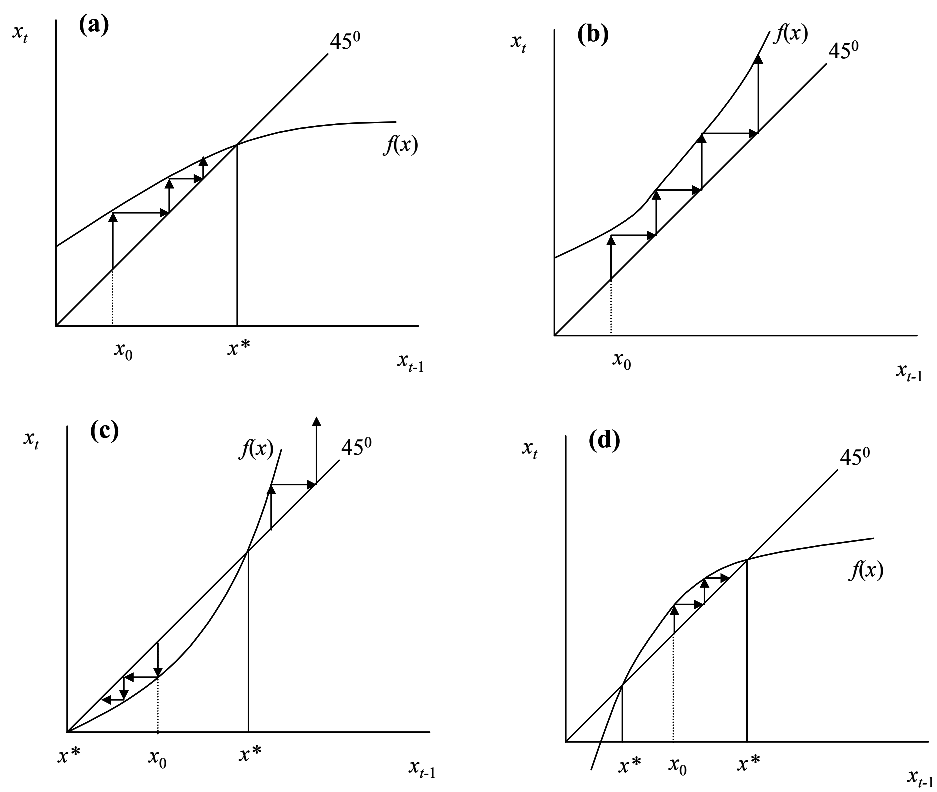


FIGURE 3.1. Phase diagrams for a nonlinear scalar equation.

Finally, in panel (c) the unique fixed point is never attained. Instead, x_t attains a limit cycle. Now, this cycle may involve the same value for x reappearing every two periods (a two-cycle), every n periods (an n-cycle), or even never (chaos). What determines the behavior of x is the shape of the function *f*. In fact, a single function may alter its behavior radically as a parameter in that function changes by a small amount, and the study of these radical changes is the essence of chaos theory.[4]

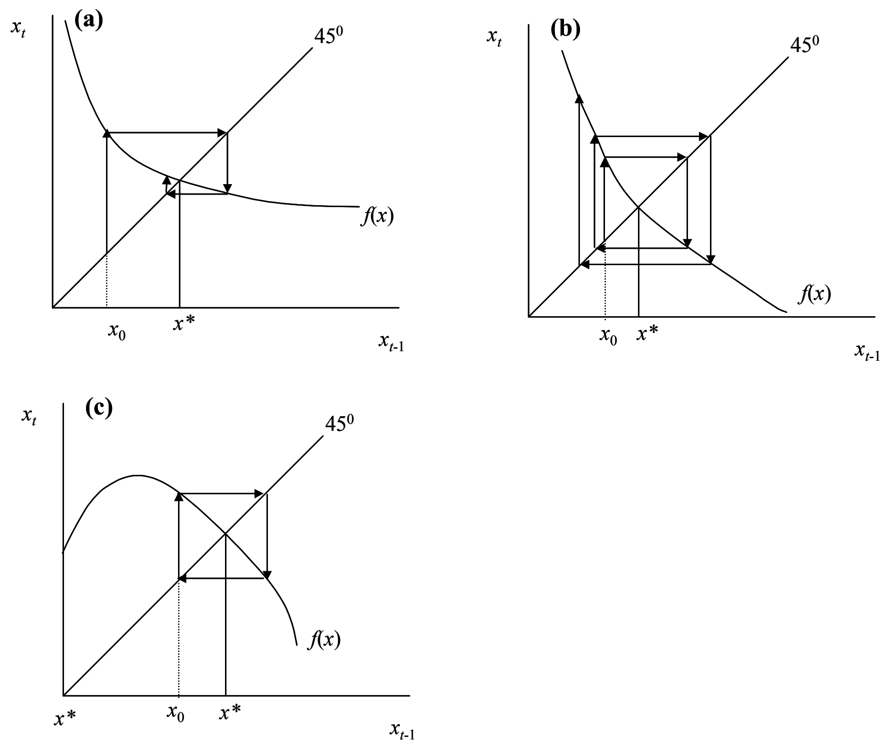
Let us return to the case of monotonically increasing functions depicted in Figure 3.1. Inspection of the phase diagrams reveal the following:

- If $f(x)$ intersects the 45^0 line with a slope less than 1, then the fixed point is stable.
- If $f(x)$ intersects the 45^0 line with a slope greater than 1, then the fixed point is unstable.
- If $f(0) \geq 0$, and $f^{/}(x) < 1$ for all x, then there is a unique fixed point.
- If $f(0) \geq 0$ and $f(x)$ is a bounded continuous function, then there must be at least one fixed point.

These intuitive statements from Figure 3.1 do indeed correspond roughly to some formal theorems, which we will explore next.

---

[4] As the parameter of a function changes, stable fixed points may suddenly become unstable, being replaced with a stable two-cycle. Stable two-cycles may in turn suddenly become unstable, being replaced with a stable four-cycle, and so on. Eventually, all fixed points and cycles of all frequency may become unstable, which is the domain of chaos. Consider the logistic equation, $x_t = bx_{t-1} - bx_{t-1}^2$. For $0<b<1$, there is a stable fixed point at x=0 and an unstable fixed point at $x=b/(b\text{-}1)$. For $1<b<3$, the larger fixed point is stable. For $b>3$, both fixed points are unstable and a stable two-cycle emerges until $b>3.45$ at which point a stable 4-cycle emerges. This process continues until $b>3.57$, beyond which all cycles of all frequencies are unstable; the value of x becomes chaotic as x wanders all over the place without any apparent repetition. See Sandefur (1990, chapter 4).



Stability of Fixed Points

FIGURE 3.2. More phase diagrams for the nonlinear scalar equation.

THEOREM 3.1. *Suppose x^* is a fixed point of the scalar dynamical model $x_t = f\left(x_{t-1}\right)$. Then x^* is attracting if $\left|f^{/}(x^*)\right| < 1$, and it is repelling if $\left|f^{/}(x^*)\right| > 1$.*

Note that Theorem 3.1 applies to all functions, monotonically increasing or decreasing, or non-monotonic. The theorem applies only to local stability. The range of values $\left\{x^* - a, x^* + b\right\}$, $a,b>0$, such that an initial value that falls in the range eventually converges on to x^* is referred to as the **basin of attraction** of the fixed point x^*.

If there is no fixed point, then we may be interested in knowing whether $x_t \to +\infty$ or $x_t \to -\infty$. The following general result provides an answer in some cases.

THEOREM 3.2. *Suppose there are no fixed points of the scalar dynamical model* $x_t = f(x_{t-1})$. *Define* $g(x_t) = f(x_t) - f(x_{t-1})$. *If* $\max g(x_t) < 0$, *then* $\lim_{t\to\infty} x_t = -\infty$. *If* $\min g(x_t) > 0$, *then* $\lim_{t\to\infty} x_t = +\infty$.

EXERCISE 3.1 (Hunting Deer). *Suppose that, in the absence of hunting, a deer population evolves according to the logistic equation* $x_t = 1.8x_{t-1} = 0.8x_{t-1}^2$, *where x is measured in thousands.*

a) *Analyze the steady state properties of this system.*

b) *Assume that permits are issued to hunt 72 deer per time period year. What happens to the steady state population? How does your answer depend on the initial population?*

c) *Assume now that permits are issued to hunt 240 deer per time period year. What happens to the steady state population? How does your answer depend on the initial population?*

Answer these questions using Theorems 3.1 and 3.2 as well as graphically.

Theorem 3.1 provides local stability results when $\left|f^{/}(x^*)\right| \neq 1$. What if $\left|f^{/}(x^*)\right| = 1$? In this case, Theorem 3.1 is unhelpful. In fact, as Figure 3.3 shows, any form of stability is possible. Although we could just look at these phase diagrams, it would be nice to have an analytical result. Fortunately, these is one, although the conditions vary depending upon whether $f^{/}(x^*) = 1$ or $f^{/}(x^*) = -1$.

THEOREM 3.3. *Suppose at a fixed point that* $f^{/}(x^*) = 1$. *Then* (a) *if* $f^{//}(x^*) < 0$, *then* x^* *is semistable from above* [*panel* (a) *of Figure 3.3*]. (b) *if* $f^{//}(x^*) > 0$, *then* x^* *is semistable from below* [*panel* (d) *of Figure 3.3*]. (c) *if* $f^{//}(x^*) = 0$, *then* x^* *is stable* *when* $f^{///}(x^{*}) < 0$ *[panel* (b) *of Figure 3.3] and unstable when* $f^{///}(x^{*}) > 0$ *[panel* (c) *of Figure 3.3].*

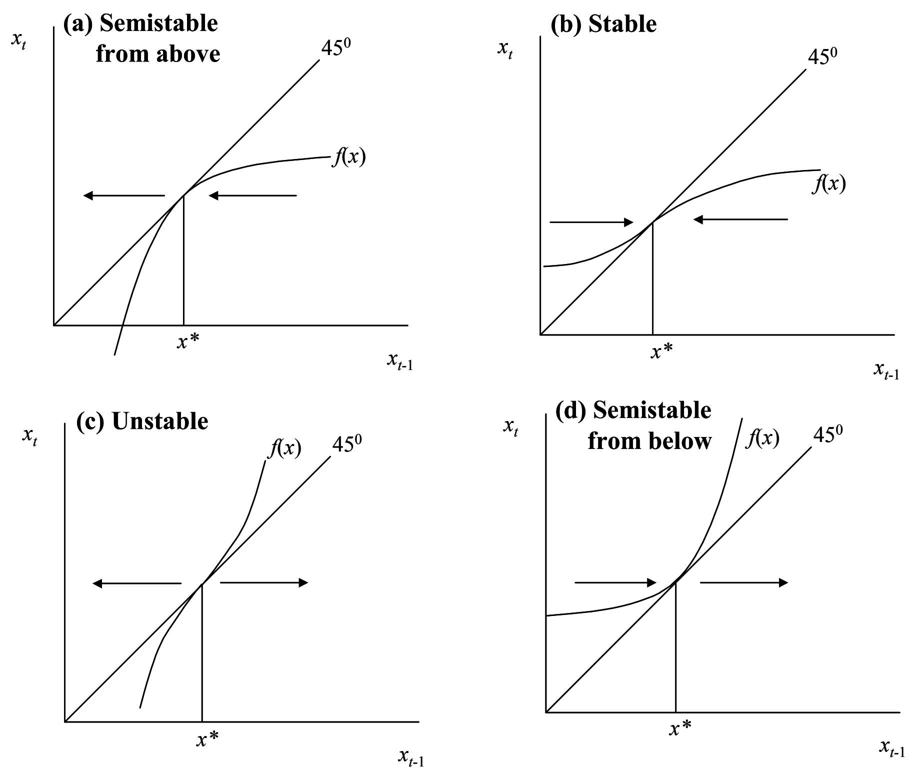


FIGURE 3.3. When $\left|f^{/}(x^{*}) = 1\right|$, all stability types are possible.

If $f^{///}(x^{*}) = 0$, results exist based on the fourth derivative. Theorem 3.3 is intuitive from Figure 3.3. The next theorem, however, is not so intuitive, but we state it without proof.

THEOREM 3.4. *Suppose at a fixed point that* $f^{/}(x^{*}) = -1$. *Define the function* $g(x^{*}) = -2f^{///}(x^{*}) - 3\left[f^{//}(x^{*})\right]^{2}$. *Then,* x^{*} *is stable when* $g(x^{*}) < 0$ *and unstable when* $g(x^{*}) > 0$.

See Sandefur (1990:165-171) for a proof.

We have yet to address the questions about existence and uniqueness of fixed points. When we have a particular functional form for $f(x)$ we can obviously address these questions directly. But, of course, in many economic applications, $f(x)$ will remain unspecified. In such cases, we have an interest in assumptions that must be made about f in order to ensure (i) the existence of at least one fixed point, or (ii) the existence of exactly one fixed point.

Existence of Fixed Points: Fixed Point Theorems for Bounded Functions

Let us begin with sufficient conditions for the existence of one or more fixed points. We make use of the following theorem:

THEOREM 3.5 (Intermediate value theorem). *Suppose $f(x)$ is a continuous function defined for any x on the compact closed interval $[a,b]$, and let c be any number satisfying $f(a) \leq c \leq f(b)$. Then there exists at least one number x such that $f(x)=c$.*

NOTE: A closed interval is one which includes its limit points. If the endpoints of the interval are finite numbers a and b, then the interval is denoted $[a,\ b]$. If one of the endpoints is ∞, then the interval still contains all of its limit points, so $[a,\infty)$ and $(-\infty,b]$ are also closed intervals. However, the infinite interval $(-\infty,\infty)$ is *not* a closed interval. A *compact* closed interval requires that a and b both be finite.

We can now state the following fixed point theorem:

THEOREM 3.6 (Fixed point theorem). *Suppose $f(x)$ is a continuous function satisfying $f(x) \in [a,b]$ for all $x \in [a,b]$* (i.e. $f(x)$ is a *mapping from the compact closed interval $[a,b]$ back into the same interval*). *Then there exists at least one fixed point $x^* \in [a,b]$.*

PROOF. Note that $f(a) \geq a$ and $f(b) \leq b$ by assumption. That is, $f(a) - a \geq 0$ and $f(b) - b \leq 0$. Define $g(x) = f(x) - x$, so that $g(b) \leq 0$ and $g(a) \geq 0$. If $f(x)$ is continuous, then so is $g(x)$. The intermediate value theorem then states that there exists at least one number x^* such that $g(x^*)=0$. This of course implies that there exists at least one number x^* such that $f(x^*)=x^*$. •

This fixed point theorem works also for n-dimensional functions, in which case it is called **Brouwer's fixed point theorem**. Yet another, the **Kakutani fixed point theorem**, also exists after the continuity assumption has been relaxed. In fact, there are numerous variations of the fixed point theorem, each with somewhat different assumptions. Many of these are given, and proved, in Border (1985). We will generally make use of the simple theorem, assuming continuity.

What does the fixed point theorem imply, and how does one make use of it? Consider the following examples:

- *In one dimension*. Simon (1954) addresses a concern that election predictions are *necessarily* self-falsifying because predictions, when they are made, influence the outcome. Now, the fraction of the vote won by a candidate is bounded between [0,1]. Thus, a prediction must come from [0,1] and, no matter how the prediction affects the outcome, the outcome must also come from [0,1]. Let y be the election outcome and x the prediction, and assume that the prediction influences the outcome according to the function $y=f(x)$. The function $f(x)$ maps a closed interval back into the same closed interval. If it is continuous, then, there must be at least one prediction that would turn out to be true.

- *In two dimensions*. Place a map of Pittsburgh (or wherever you happen to be) on the floor. The fixed point theorem states that at last one point on the map lies exactly above its corresponding point in the city. The reason is that one can define a two-dimensional function f linking every location in the city, (x,y) to another location in the city, $f(x,y)$ (i.e. the floor under the map) that touches the corresponding location on the map. This function is continuous and bounded (by the size of the city). Thus there is at least one fixed point where $(x,y)=f(x,y)$.

- *In three dimensions*. Take your cup of coffee and swirl it around. Let (x,y,z) denote the coordinates of a molecule of coffee before you start swirling, and let $f(x,y,z)$ denote its location after. Because the movement of liquid so agitated can be described by a system of differential equations, the function f must be continuous. It is also bounded by the coordinates of the edge of the cup. Thus, unless you spill some coffee on the floor (in which case f is no longer mapping a closed interval back into the same interval), there must be at least one molecule in exactly the same position after swirling as before.

OK, these are just amusing examples, but they give a straightforward flavor of the meaning of the fixed point theorem. How one uses the theorem in practice is also straightforward in principle. Given a function $y=f(x)$, we verify whether it satisfies the assumptions of Theorem 3.6 (or one of the more general theorems to be found in Border [1985]). If so, then we know there is at least one fixed point.

Note that each assumption is critically important. Consider, for example, the discrete-time Solow model in which capital evolves according to the nonlinear difference equation

$$k_t = sf(k_{t-1}) - (1-\delta)k_{t-1}\,, \tag{3.3}$$

Unless we make some assumptions about the production function, this difference equation need not have a fixed point. In particular, if f is a convex increasing function the right hand side of (3.3) does not map a closed compact interval into a closed compact interval. Thus the fixed point theorem cannot be applied, even though one may exist. For example, if the production function has the form $f(k) = k^2$, then a fixed point exists at k=0. In contrast, if $f(k) = 1 + k^2$, then no fixed point exists. One way to ensure the existence of a fixed point is to place an upper bound on how much a nation can produce regardless of its capital stock. That is $f(k) < \overline{f}$ for all k. Then, k_t can never exceed an upper bound, say $\overline{k}$, and (3.3) defines a mapping from the closed compact interval $\left[0,\overline{k}\right]$ back into $\left[0,\overline{k}\right]$. Now, the fixed point theorem applies and there exists at least one fixed point.[5]

As Figure 3.4 shows, the assumption of continuity is also an important one (and theorems such as Kakutani's can relax it only in very precise ways). If the function $f(x)$ is not continuous, then there need not be a fixed point. Thus, the challenge in Simon's election example, is justifying on behavioral grounds that the function f is indeed continuous. The challenge in the coffee swirling example is also showing that the physics of movement of liquids imply that f is continuous.

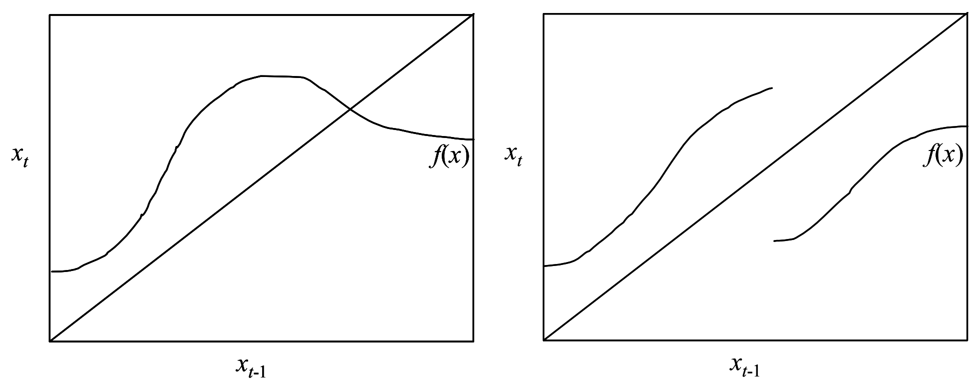


FIGURE 3.4. Left Panel: a bounded continuous function must cross the 45 degree line at least once. Right Panel: a bounded discontinuous function need not have a fixed point.

Of course, the theorem cannot tell you where these fixed points are. But finding fixed points can be quite difficult. And if you are going to look for a needle in a haystack, it is good to know that there are some needles there before you start.

[5] Recall that the Inada conditions studied in the chapter on differential equations, which placed boundary values on the derivatives of the production function, were designed to ensure the existence of a fixed point.

EXAMPLE 3.2. Is there a real solution to the equation

$$\frac{\sqrt{x}}{1-e^{-3x}} = 1\,? \tag{3.4}$$

Clearly, x must be nonnegative for a real solution to exist, so we restrict x to be nonnegative. Rearrange the equation to give

$$x = \left(1-e^{-3x}\right)^2 . \tag{3.5}$$

The right hand side is bounded between 0 and 1, and it is clearly continuous (you can differentiate it!). Hence there is at least one fixed point in the unit interval. Note also that we can write this as a difference equation

$$x_t = \left(1-e^{-3x_{t-1}}\right)^2 , \tag{3.6}$$

which makes clear that solving (3.5) is equivalent to finding the fixed point(s) of a difference equation. Doing so may give us a way to numerically solve the equation. Simply begin with an initial value in the unit interval and iterate the difference equations to convergence. If it converges, a fixed point has been found.

We report in the table below the iterations for three different starting values. In the first column, a starting value of 0.19 found a fixed point x^*=0. In the third column, a starting value of 0.99 yielded a fixed point of x^*=0.8499. Convergence had not quite been attained by iteration 15 in the second column, which had a starting value of 0.20, but it eventually found the fixed point x^*=0.8499. The reason for the behavior of these iterations will become quite clear upon plotting the phase diagram for (3.6). There are two stable fixed points, at x^*=0 and x^*=0.8499, and one unstable fixed point at x^*=0.193.

The numerical method is not very efficient, and it cannot find unstable fixed points. But the basic principle of numerical solution of implicit equations is nonetheless quite clear. However, this particular example also reveals a hidden danger. Equation (3.5) has a fixed point x^*=0. Yet x=0 is not a solution to (3.4). To see this, note that by l'Hôpital's rule,

$$\lim_{x\to 0}\frac{\sqrt{x}}{1-e^{-3x}} = \lim_{x\to 0}\frac{\frac{1}{2}x^{-1/2}}{3e^{-3x}} = +\infty \neq 1 .$$

Iterated Values for $f(x)$

	INITIAL VALUE		
ITERATION	0.1900	0.2000	0.9900
1	0.1888	0.2036	0.9000
2	0.1870	0.2089	0.8701
3	0.1843	0.2168	0.8584
4	0.1804	0.2268	0.8535
5	0.1746	0.2464	0.8514
6	0.1663	0.2730	0.8506
7	0.1543	0.3126	0.8502
8	0.1373	0.3703	0.8500
9	0.1139	0.4499	0.8499
10	0.0838	0.5486	0.8499
11	0.0494	0.6515	0.8499
12	0.0190	0.7638	0.8499
13	0.0031	0.7927	0.8499
14	0.0001	0.8231	0.8499
15	0.0000	0.8379	0.8499

In contrast, x=0 *is* a solution to

$$x = \left(1 - e^{-3x}\right)^2,$$

which shows that you have to be careful about manipulating equations that may take on the value 0/0. Curious. •

Before closing this subsection, note that the fixed point theorem for bounded functions is a sufficient condition, not a necessary one. Unbounded functions may have fixed points, but the fixed point theorem cannot help you there – you have evaluate the function yourself. For example, the difference equation $x_t = 1 + 0.5x_{t-1}$ has a fixed point at x*=2, but it is not bounded. The function

$$x_t = \begin{cases} 1, & x_{t-1} \le 2 \\ 0.5, & x_{t-1} > 2 \end{cases}, \tag{3.7}$$

has a fixed point at x^*=1, but it is not continuous. However, the function (3.7) defined only on the closed compact interval $x \in [0,2]$ *is* a continuous mapping from [0,2] back into [0,2], and the fixed point theorem does apply.

Uniqueness of Fixed Points

If you are looking for a fixed point of an equation – whether it is because you are trying to find the long-run behavior of a dynamic system or because you are trying to find the solution of an equation defined only implicitly – you would also like to know if there is only one needle in the haystack. Once you find one, you can stop looking if you know there are no more. Fortunately there is a theorem – the **contraction mapping theorem** – that provides the conditions under which the fixed point is unique.

THEOREM 3.7 (Contraction mapping theorem). *Define a function $f(x)$ which maps a value $x \in [a,b]$ into the same bounded closed interval $[a,b]$. If $f(x)$ is a contraction mapping, then there exists exactly one fixed point $x^* = f(x^*)$.*

If we can show that $f(x)$ is a contraction mapping, then we have proved not only that there is a fixed point, but also that it is unique. But what the hell *is* a contraction mapping, and how do we know if we have one? The full definition of a contraction mapping requires that we introduce a bunch of new concepts – Cauchy sequences, convergent sequences, metric spaces, complete metric spaces, and more -- that build one on the other like a set of Russian dolls. We will leave as much of that as possible to the theorists (and the budding theorists should refer to Stokey and Lucas [1989]). What we will do here is provide an intuitive definition of a contraction mapping.

DEFINITION (Contraction mapping): *Let $f(x)$ be a function mapping elements from the closed interval $[a,b]$ back into the same closed interval. Let $\{x,y\}$ be two real numbers from the closed interval, and let $d(x,y)$ be a distance function that measures the distance between*

x and y in an appropriate sense to be discussed below. Then, the function $f(x)$ is a contraction mapping if $d\left(f(x), f(y)\right) < d(x, y)$ for any x and y from the closed interval.

Intuitively, f is a contraction mapping if operating on x and y with the function f makes them closer together. However, we do need to deal with the notion of an "appropriate" measure of distance. For a function $d(x,y)$ to be an admissible distance function, it must satisfy the following properties:

- *Nonnegativity*: $d(x,y)>0$ if $x \neq y$, and $d(x,y)=0$ if $x=y$.
- *Symmetry:* $d(x,y)=d(y,x)$.
- *Triangle inequality:* $d(x,y)+d(y,z)\geq d(x,z)$.

The first condition states that the distance between x and y is positive if they are not equal and zero if they are. The symmetry property states that x is the same distance from y as y is from x. For the third property, let x be Pittsburgh, y Philadelphia, and z Washington. The triangle inequality states that the distance between Pittsburgh and Washington via Philadelphia cannot be less than the direct distance between Pittsburgh and Washington. These are reasonable properties to demand of a function intended to measure distance, and they delimit the sort of functions that that are admissible. For example, the function $d(x, y) = x - y$ is not admissible, because it violates at least the first two properties. However, the functions $d(x, y) = |x - y|$ and $d(x, y) = (x - y)^2$ do satisfy the three properties, so they are admissible distance functions.

Now we have defined a contraction mapping, we can prove the uniqueness result in the contraction mapping theorem. It is easy. Given a difference equation $x_t = f(x_{t-1})$, suppose that f is a contraction mapping, and suppose there were two distinct fixed points, x^* and x^{**}. Then for any distance function d, $d(f(x^*),f(x^{**}))<d(x^*,x^{**})$, since f is a contraction mapping. But as these are fixed points, $x^*=f(x^*)$ and $x^{**}=f(x^{**})$, so $d(x^*,x^{**})=d(f(x^*),f(x^{**}))<d(x^*,x^{**})$, a contradiction. Thus, there cannot be two fixed points. •

The requirement that $d\big(f(x), f(y)\big) < d(x, y)$ for any x and y effectively means that $f(x)$ must be a continuous function. To see why, consider the right-hand panel of Figure 3.4. Clearly, if the function $f(x)$ is not continuous, $d\big(f(x), f(y)\big)$ cannot be less than $d(x,y)$ for values of x and y close enough to, and opposite sides of, the discontinuity. Second, for *differentiable* functions (although the contraction mapping theorem does not require differentiability), it is easy to see that, if $d\big(f(x), f(y)\big) < d(x, y)$ for all x and y, then $f^{/}(\cdot)$ must be less than one in absolute value everywhere.

We are almost done now, so this is a good time to summarize. Imagine you have a difference equation, $x_t = f(x_{t-1})$. You want to verify whether this difference equation has any fixed points, and whether it has a unique fixed point. There are several steps to go through.

(1). Is the function a mapping from a compact closed interval back into the *same* compact closed interval? Remember that a compact closed interval means that the interval includes its own finite end points.

(2). If (1) is satisfied, then is the function $f(x)$ continuous? If it is, then you have established that at least one fixed point exists. If not, then turn to Border (1985) to see if the function satisfies a more general fixed point theorem. If Border does not help you, then you cannot conclude anything about whether there is a fixed point, and you have to get to work to find some yourself (without knowing in advance whether there are any).

(3). We next turn to the question of uniqueness. We need to check whether the function $f(x)$ is a contraction mapping by checking that an appropriate distance function satisfies $d\big(f(x), f(y)\big) < d(x, y)$ for any feasible pair of values, x and y.

This last step is invariably the most difficult one. Fortunately, as we have already noted a bit of calculus can help us for differentiable functions because if $\left|f^{/}(x)\right| < 1$ for all $x \in [a, b]$, then we have uniqueness. This is easy to see if we let the absolute difference between x and y be our distance function:

$$\left|f(x)-f(y)\right| = \left|\int_x^y f^{/}(s)ds\right| < \left|\int_x^y ds\right| = \left|y-x\right|,$$

so we have a contraction mapping as long as the function f is a mapping in a closed interval.

EXAMPLE 3.3. Let $x \in [0,1]$, and consider the difference equation $x_t = f(x_{t-1})$. Check the following functions for existence and uniqueness of fixed points.

(a) $f(x) = \ln(x+0.5)+0.2$. This fails the first step because it does not map into the unit interval. For example, for x=0, $f(x)$=–0.4931. Nor can we adjust the closed compact interval to accommodate the negative number: choose any negative number a for the lower bound on the interval and you will find that $f(a)<a$ because $f^{/}(x)>1$ for any $x<0$. This also implies, of course, that the function is not a contraction mapping.

(b) $f(x) = x+0.5$ for $x<0.5$, and $f(x) = x-0.5$ for $x\geq0.5$. This passes the first test, as it maps [0,1] back into [0,1] but it fails the continuity test.

(c) $f(x) = \ln(x+1)+0.2$. This passes the first test, as it maps [0,1] back into the same interval (in fact $f(x)$ ranges between 0.2 and 0.89). It passes the continuity test in this interval also.[6] Thus there is at least one fixed point. To check for uniqueness we need to evaluate whether $\left|(\ln(x+1)+0.2)-(\ln(y+1)+0.2)\right| < \left|x-y\right|$. As this is difficult to do directly, we note that $f^{/}(x) = 1/(x+1)$ which is less than one in absolute value almost everywhere (i.e. except when x=0). Thus, there is a unique fixed

[6] Continuity is often checked by inspection of the function. However, to see how it might be done formally, note that $f(x)$=1/(x–1) does not pass a continuity test in the range [0,2]. This is formally shown by identifying a value of x, x=z, such that the value of the function as x approaches z from below is different from the value of the function as x approaches z from above. In this case, as x approaches 1 from below, the function tends to $-\infty$; as it approaches 1 from above the function tends to $+\infty$. However this function *does* have a fixed point at x*=1.618. In contrast, the function $f(x)$=1/(1–x) also has a discontinuity at x=1, but it does not have a fixed point anywhere.

point. Moreover, Theorem 3.1 tells us that in this case the fixed point is attracting. •

(d) Does the difference equation $x_{t+1} = 3 + 2x_t$, defined on a closed interval, have a unique fixed point? That is, is it a contraction mapping? We can see immediately that $f^{/}(x) = 2$, so the contraction mapping theorem does not work. However, we can *invert* the equation and write it as $x_t = 0.5x_{t+1} - 1.5$, which now is a contraction mapping. Hence there is a unique fixed point. It turns out that this fixed point is at $-x^*$. Whether this makes sense to do depends on whether the reversibility of the time subscript is economically meaningful. If there is a notion of causation, then it might not be. If this is a means to find a fixed point of an implicit equation, then it does make sense.

This must all seem a little convoluted. Why not just study the differential equation directly? For many applications that is exactly what one should do. But recall that in many economic models we want to make as few assumptions about functional form as possible, and then see what predictions we can make. But how can we solve for the fixed point without specifying the function *f*? The answer is we do not have to. Consider the following example. Let s_t denote a firm's market share and let the evolution of market share satisfy the difference equation $s_t = f(s_{t-1};\beta)$, where β is a parameter. Because market shares are bounded, *f* must be a mapping from [0,1] into [0,1]. If we are willing to assume that *f* is a continuous function, then we know there is at least one fixed point. Let us further assume that $0 < f_s(s;\beta) < 1$, where the subscript denotes the derivative. That is, if there is an autonomous shock to market share this period, only a fraction of this shock persists into the next period. This assumption that shocks to market share decay means that *f* is a contraction mapping, so now we have established that there is a unique fixed point. Moreover, Theorem 3.1 tells us this unique fixed point is stable. Having made this much progress, it is now an easy matter to see how the long-run market share of the firm is affected by changes in the parameter. Differentiating at the unique fixed point gives $ds^* = f_s(s^*;\beta)ds^* + f_\beta(s^*;\beta)d\beta$, or

$$\frac{ds^*}{d\beta} = \frac{f_\beta(s^*;\beta)}{1 - f_s(s^*;\beta)}. \tag{3.8}$$

That is, having a theory about how changes in the parameter β affect market share at any point in time allows us to derive an expression for changes in the steady-state market share without ever actually solving for the steady state. Equation (3.8) tells us that the total effect of a change in β on the long-run market share consists of two terms. The first is the direct effect, f_β, while the second is the feedback effects $1/(1-f_s)$. Note also why we would like to have a unique fixed point for this analysis (but whether we have one depends on the economics of the problem). If s^* is unique, then (3.8) provides a unique answer to the question of how parameter changes affect long-run market share. If there are multiple fixed points, the answer is different at each one.

Notes on Further Reading

References

Bergstrom, A.R. (1990): *Continuous Time Econometric Modelling*. Oxford: Oxford University Press.

Border, Kim C. (1985): *Fixed Point Theorems with Applications to Economics and Game Theory*. Cambridge, UK: Cambridge University Press.

McCallum, Bennett T. (1983): "On Non-Uniqueness in Rational Expectations Models: An Attempt at Perspective." *Journal of Monetary Economics*, 11:139-168.

Sandefur, James T. (1990): *Discrete Dynamical Systems*. Oxford: Oxford University Press.

Sargent, Thomas J. (1979): *Macroeconomic Theory*. New York: Academic Press.

Simon, Herbert (1954): "Bandwagon and underdog effects and the possibility of election predictions." *Public Opinion Quarterly*, 18:245-253.

Stokey, Nancy L., and Robert E. Lucas Jr. (1989): *Recursive Methods in Economic Dynamics*. Cambridge, MA: Harvard University Press

Taylor, John (1979): "Staggered Wage Setting in a Macro Model", *American Economic Review Papers and Proceedings*, 69:108-113