Impact Project

Impact Centre
The University of Melbourne
153 Barry Street, Carlton
Vic. 3053 Australia

Phones: (03) 341 7417/8
Telex: AA 35185 UNIMEL
Telegrams: UNIMELB, Parkville

IMPACT is an economic and demographic research project conducted by Commonwealth Government agencies in association with the Faculty of Economics and Commerce at The University of Melbourne and the School of Economics at La Trobe University.

NUMERICAL METHODS FOR INVESTMENT MODELS WITH FORESIGHT

by

Peter J. Wilcoxen
University of Melbourne
and
Harvard University

Preliminary Working Paper No. IP-23 Melbourne July 1985

ISBN 0 642 52650 8

This paper is also issued as

Research Paper No.132

of the Department of Economics

University of Melbourne

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

The author would like to thank Alan Powell for helpful comments on an earlier draft of this paper, and the University of Melbourne for its financial support.

NUMERICAL METHODS FOR INVESTMENT MODELS WITH FORESIGHT

by

Peter J. Wilcoxen

1. Introduction

Economic models of foresight typically require the solution to an intertemporal maximization problem. Finding the first-order conditions is usually straightforward, but it is often impossible to find an explicit expression for the optimal path these conditions describe. Fortunately, such equations can be solved using numerical analysis, although economic problems possess a number of properties which make this difficult. This paper discusses some of these characteristics and presents a versatile method for finding a numerical solution to the first-order conditions. It will be shown that the technique, known as finite differences, is easy to apply and provides accurate results at low cost.

To illustrate the use of numerical analysis, a typical foresight model will be developed in detail. The economic features of this particular model are discussed in more detail elsewhere (Wilcoxen (1985), hereafter referred to as the descriptive paper); attention here will be limited to the aspects of the problem which are important for numerical analysis. This sample problem is the firm's choice of investment to maximize its market value, but the techniques used are applicable to any foresight model which generates first-order conditions of similar structure.

1.1 Notational Conventions

The only unusual notation used here is that the derivative of a variable

with respect to time is indicated by the prime (′) symbol, rather than the usual dot.

2. The Constant Returns to Scale Investment Model

A familiar case in which foresight plays an important role is the firm's problem of choosing investment to maximize its stock market value. Given a short run earnings function, E(K), and an investment cost function C(I), the firm must solve:

$$\max \int_t^\infty \{ E(K)-C(I) \}(1-T^d)\exp(-rs)ds ,$$

$$\text{subject to} \quad K' = I - \delta K ,$$

where r is the interest rate, δ is the rate of depreciation, and T^d is the dividend tax rate. Applying the principles of optimal control (discussed in the descriptive paper) gives the following first-order conditions:

$$\lambda = \frac{\partial C}{\partial I}(1-T^d) ,$$

$$\lambda' = (r+\delta)\lambda - \frac{\partial E}{\partial K}(1-T^d) ,$$

$$K' = I - \delta K ,$$

where λ is the current value multiplier associated with the constraint on the derivative of capital. As an example, suppose the earnings and investment cost functions are:

$$E(K) = \beta K ,$$

$$C(I) = P_k\left((1+\gamma)I + \theta I^2\right)(1-T^s) ,$$

where β, γ and θ are constants, with θ>0; T^s is an investment subsidy; and P_k is the price of new capital goods. The investment cost function is convex in

References

Birkhoff, G. and G-C Rota: Ordinary Differential Equations. London: Blaisdell, 1969.

Dixon, P.B., et al.: ORANI: A Multisectoral Model of the Australian Economy. Amsterdam: North-Holland, 1982.

Fox, L.: Numerical Solution of Ordinary and Partial Differential Equations. London: Addison-Wesley, 1962.

Isaacson, E. and H.B. Keller: Analysis of Numerical Methods. New York: John Wiley and Sons, 1966.

Leitmann, G.: The Calculus of Variations and Optimal Control: An Introduction. New York: Plenum Press, 1981.

Lipton, D., et al.: "Multiple Shooting in Rational Expectations Models," Econometrica, 50(1982), 1329-1333.

Roberts, S.M. and J.S. Shipman: Two-Point Boundary Value Problems: Shooting Methods. New York: American Elsevier, 1972.

Wilcoxen, P.J.: "Computable Models of Investment with Foresight," mimeo, University of Melbourne, April, 1985.

where a and b are small positive increments in time. Unfortunately, accuracy of these formulae is diminished at points where there are sharp changes in step size; to illustrate this, consider the Taylor expansion of the first-order difference above:

$$f(t+a) - f(t-b) =$$
$$\left(f(t) + af'(t) + a^2f''(t)/2! + a^3f'''(t)/3! + O(a^4) \right) -$$
$$\left(f(t) - bf'(t) + b^2f''(t)/2! - b^3f'''(t)/3! + O(b^4) \right) ,$$

so cancelling terms and dividing by a+b gives the difference formula:

$$\frac{f(t+a) - f(t-b)}{a+b} = f'(t) + (a-b)f''(t)/2! + \left(\frac{a^3+b^3}{a+b}\right)f'''(t)/3! + \ldots$$

Thus grid nonuniformity introduces an additional source of truncation error into the solution through terms in (a-b). This effect can be minimized by avoiding sharp jumps in grid spacing, and by locating any such changes in regions where f'' is small. Further, by constructing more elaborate difference formulae, the term in (a-b) can be eliminated entirely. For additional discussion of difference formulae in general, consult Fox (1962).

investment, so the firm faces diminishing returns when adding new capital (refer to the descriptive paper for further details).

Inserting these functions into the general first-order conditions gives the following:

$$\lambda = P_k(1+\gamma+2\theta I)(1-T^d)(1-T^s) ,$$

$$\lambda' = (r+\delta)\lambda - \beta(1-T^d) ,$$

$$K' = I - \delta K .$$

These equations may be rewritten as a single second-order equation in capital, as shown below:

$$K'' + K'(g-r) + \delta K(g-r-\delta) = \frac{1}{2\theta}\left((1+\gamma)(r+\delta-g) - \frac{\beta}{P_k(1-T^s)}\right) ,$$

where

$$g = -\frac{T^{s\prime}}{1-T^s} - \frac{T^{d\prime}}{1-T^d} .$$

The second-order equation describes a two-parameter family of optimal paths of the capital stock; a complete solution requires that these parameters, which arise as constants of integration, be specified. In an "initial value" problem the constants will be chosen to make the system meet conditions which apply to the initial point; for example:

$$K(0) = Ko ,$$

$$K'(0) = Ko' ,$$

where Ko and Ko' are particular values of the initial capital stock and its derivative. The second-order equation and the initial conditions together describe the optimal path of the capital stock.

In economic problems, however, the value of Ko′ is usually unknown, so it is necessary to specify the second boundary condition differently. In place of this missing initial condition the model can be required to approach its steady state as time goes to infinity; technically, this is known as a transversality condition. The boundary conditions then become:

$$K(0) = Ko \ ,$$

$$\lim_{t\to\infty} \quad K(t) = K^{ss} \ .$$

When the conditions are specified in this way the system is referred to as a "two-point" boundary value problem, as the boundary conditions are specified at separate points in time. Finding the actual path of the capital stock from the first-order conditions for value maximization thus requires solving a two-point boundary value problem. Unfortunately, it will often be impossible to find an analytic solution to this problem, and numerical methods will have to be used. Such techniques are the primary topic of this paper and will be discussed in the next section. First, however, it is instructive to examine a simple example which can be solved analytically.

It is easy to find a solution to model above when the dividend tax and investment subsidy are expected to be constant forever. In that case the problem becomes:

$$K'' - rK' - \delta(r+\delta)K = \frac{1}{2\theta}\left((1+\gamma)(r+\delta) - \frac{\beta}{P_k(1-T^s)}\right) \ .$$

This is a non-homogenous linear (in K) second-order differential equation, and it can be solved by the usual technique of finding a solution to the related homogenous problem and adding a "particular" solution. The homogenous solution to the equation above is:

Appendix: Construction of Finite Difference Formulae

The finite difference formulae used here, technically known as divided differences, are constructed by combining the Taylor series expansions of a number of adjoining grid points. For example, the first order central divided difference is obtained by subtracting the expansion for f(t-h) from that for f(t+h) and dividing by 2h, as shown below:

$$\begin{aligned} f(t+h) - f(t-h) = \\ &(f(t) + hf'(t) + h^2f''(t)/2! + h^3f'''(t)/3! + O(h^4)) - \\ &(f(t) - hf'(t) + h^2f''(t)/2! - h^3f'''(t)/3! + O(h^4)) \ . \end{aligned}$$

Cancelling out terms gives the following:

$$f(t+h) - f(t-h) = 2hf'(t) + 2h^3f'''(t)/3! + O(h^5) \ ,$$

so the central divided difference formula approximates the derivative of f(t) to $O(h^2)$:

$$f'(t) = \frac{f(t+h) - f(t-h)}{2h} + O(h^2) \ .$$

Since the low-order error term is $h^2f'''/3!$, truncation error will generally be largest where f''' is large.

A similar approach is used when constructing formulae for nonuniform grids and the particular approximations used in the experiments presented here are those shown below:

$$f'(t) \cong \frac{f(t+a) - f(t-b)}{a+b} \ ,$$

and

$$f''(t) \cong \frac{f(t+a)}{a(a+b)/2} - \frac{2f(t)}{ab} + \frac{f(t-b)}{b(a+b)/2} \ ,$$

7. Conclusion

Solving most economic models of foresight will produce a system of differential equations and a set of boundary conditions which together provide a complete, although implicit, description of the optimum. Finding an explicit expression for the solution will be complicated by the inherent difficulty of solving systems of differential equations, and by the form of the usual boundary conditions. In fact, it will be impossible to solve most models analytically and numerical methods will have to be used.

The results presented here indicate that techniques are available which produce accurate results with a minimum amount of problem preparation and at minimum computational cost. The finite difference method in particular is well suited to economic problems because of its numerical stability. Furthermore, it is highly compatible with the solution algorithms used for most general equilibrium models. Unfortunately, since an n-period foresight model requires adding n equations to the model, it may be impossible to model each sector's investment independently in a highly disaggregated model, but overall the analysis indicates that no substantial difficulty exists which would prevent construction of general equilibrium models which have some degree of foresight.

$$K_h = C^1\exp((r+\delta)t) + C^2\exp(-\delta t) ,$$

and one "particular" solution can be found by setting $K'' = K' = 0$ in the second-order equation to get:

$$K_p = K^{ss} = \frac{1}{2\delta\theta}\left(\frac{\beta}{P_k(r+\delta)(1-T^s)} - (1+\gamma) \right) ,$$

which is also the steady state capital stock. The general solution is, therefore:

$$K(t) = C^1\exp((r+\delta)t) + C^2\exp(-\delta t) + K^{ss} , \qquad (2.1)$$

where the constants are to be chosen to satisfy the boundary conditions. Imposing those conditions gives two simultaneous equations for C^1 and C^2:

$$K(0) = Ko = C^1 + C^2 + K^{ss} ,$$

$$\lim_{t\to\infty} K(t) = K^{ss} = \lim_{t\to\infty} C^1\exp((r+\delta)t) + \lim_{t\to\infty} C^2\exp(-\delta t) + K^{ss} .$$

The second condition will only be satisfied when $C^1=0$, so the first equation requires that $C^2=Ko-K^{ss}$. The complete solution is thus:

$$K(t) = K^{ss} - (K^{ss}-Ko)\exp(-\delta t) .$$

Unfortunately, it is impossible to find an analytic solution to the general problem where the tax and subsidy are functions of time. It is possible, however, to solve the model numerically, and several of the available methods are discussed in the following section.

3. Numerical Methods - A Brief Survey

When the model can be formulated as an initial value problem, solving it requires nothing more than integrating the differential equation forward from the initial conditions. In a two-point problem, however, one condition is specified at the terminal point, so simple integration is impossible and more sophisticated techniques must be used. A large number of methods have been developed to solve such problems, but those considered here will be limited to shooting, multiple shooting and finite differences.

3.1 Shooting

One obvious way to solve the two-point problem is to guess the value of the missing initial condition, integrate forward and check whether the terminal condition is satisfied. If it isn't, the guessed condition is revised and the process repeated. This is the basic approach used in simple shooting algorithms. A model can be thought of as an implicit function that gives a "miss distance" (the difference between the computed value at the terminal time and the true terminal condition) as a function of the guessed initial conditions. An iterative technique such as Newton's Method can be applied to find the initial condition which sets the miss distance to zero. For a single variable K required to attain boundary value K(T) at time T, the miss distance is given by:

$$M(\tilde{K}o') = K_T - \kappa(0,T,Ko,\tilde{K}o') ,$$

where $\kappa(0,T,Ko,\tilde{K}o')$ is the value of K at time T found by integrating the first-order conditions forward from time zero using initial conditions Ko and $\tilde{K}o'$ ($\tilde{K}o'$ is the guessed initial condition). A first-order Taylor series expansion of M about $\tilde{K}o'$ for a trial solution $\hat{K}o'$ gives:

In contrast, if the foresight model is to be added to a general equilibrium system, this coefficient nonlinearity is important because some of the variables exogenous to the firm will be endogenous in the full system and hence not independent of K; for example, the price of capital. Fortunately, this form of nonlinearity will not have any effect on the solution procedure used for the general equilibrium model. To see why this is so consider adding a model which is linear in K to a larger system. If the composite model is solved by linearization (Johansen's Method, etc.), coefficient nonlinearity is handled in the same way as nonlinearity in other variables. On the other hand, if an iterative technique such as Scarf's Algorithm is used, no difficulty is created by coefficient nonlinearity because the foresight model is simply solved as if the current price vector were correct; the coefficients can be evaluated by direct computation using the guessed prices, so the path of the capital stock can be found by elimination at each iteration. (Nonlinearity in the differential equations sense (in K) is important, however, for models solved by iterative techniques because the foresight submodel will have to be solved by Newton's Method for the current price vector at each iteration in the overall solution.)

Overall, coefficient nonlinearity is unimportant in partial equilibrium analysis and it does not make necessary any fundamental changes in solution algorithms for general equilibrium models. It will often be desirable, however, to linearize models which are nonlinear in the differential equations sense, particularly if the foresight model is part of a general equilibrium model solved by iteration.

6. Linearity

Linearity of a foresight model is important in two distinct senses and it is worthwhile to discuss both briefly in the context of the constant returns model. First, in the terminology of differential equations, the model is said to be linear because it can be written in the form shown below:

$$a(t)K'' + b(t)K' + c(t)K = d(t) .$$

In contrast, a model of the form:

$$a(t)K'' + b(t)K' + c(t)K^2 = d(t) ,$$

is not linear. As noted above, linearity of the original model (in the differential equations sense) means that the finite difference form will be a system of linear equations which can be solved by simple elimination. If the model is nonlinear, however, the system must linearized or solved using an iterative procedure such as Newton's Method. Techniques for nonlinear equations are considerably more cumbersome (and expensive) than elimination, so in many cases it will be worthwhile to linearize the model, even at the cost of introducing a further source of truncation error.

The second sense in which the model may be linear is that the coefficients, for example d(t), may be linear functions of model variables such as the price of capital. In general, foresight models will not be linear in this sense; the coefficients will be nonlinear functions of variables exogenous to the firm's investment decision. In a partial equilibrium analysis, however, the time path of each exogenous variable will be known prior to computing the solution to the firm's problem, so nonlinearity of the coefficient functions is unimportant; each coefficient can be evaluated at each grid point by direct computation.

$$M(\hat{K}o') = M(\tilde{K}o') + M'(\tilde{K}o')(\hat{K}o' - \tilde{K}o') ,$$

which, setting $M(\hat{K}o')$ to zero since $\hat{K}o'$ is a solution, can be written as:

$$\hat{K}o' = \tilde{K}o' - \frac{M(\tilde{K}o')}{M'(\tilde{K}o')} .$$

This can be used to update the guess of Ko'. The derivative of the miss distance with respect to the initial condition is typically found by numerical differentiation.

3.1.1 Numerical Instability

Unfortunately, in economic applications the simple shooting algorithm frequently fails to converge. This is partly because Newton's Method is a local approximation and converges only when the initial guess is sufficiently close to the true solution. A more important difficulty, however, is that economic problems are often very sensitive to the value of the missing initial condition because of the presence of exponential growth terms in the solution. This problem can be illustrated by considering how the analytic solution to the constant returns model from section 2 responds to small perturbations of the boundary conditions.

The general solution to the model when all taxes are expected to be constant in the future was given in equation (2.1), repeated below:

$$K(t) = C^1\exp((r+\delta)t) + C^2\exp(-\delta t) + K^{ss} .$$

In the two-point form of the problem, the transversality condition requires that $C^1=0$, which implies that $C^2=Ko-K^{ss}$; shooting is, in essence, a method of finding an initial value problem which generates an equivalent solution. In

this related problem, boundary conditions such as the following will be given:

$$K(0) = Ko ,$$

$$K'(0) = Ko' .$$

These conditions are imposed by differentiating the general solution and evaluating both $K(0)$ and $K'(0)$ to give the following:

$$Ko = C^1 + C^2 + K^{ss} ,$$

$$Ko' = (r+\delta)C^1 - \delta C^2 .$$

Solving these for C^1 and C^2 and substituting into the general solution gives the equation below which describes the path of the capital stock for given values of Ko and Ko':

$$K(t) = \left(\frac{Ko'+\delta(Ko-K^{ss})}{r+2\delta}\right)\exp((r+\delta)t) - \left(\frac{Ko'-(r+\delta)(Ko-K^{ss})}{r+2\delta}\right)\exp(-\delta t) + K^{ss} .$$

In the solution to the two point problem, as $t\to\infty$, $K(t)\to K^{ss}$ so the equivalent initial condition is $Ko'=-\delta(Ko-K^{ss})$ and the coefficient of the positive exponential term is zero. Now consider what happens during shooting: a value of Ko' is guessed and K(t) is calculated for some large t. If the guess for Ko' is slightly incorrect, the coefficient of the positive exponential term will be non-zero and the miss distance will grow in proportion to $\exp((r+\delta)t)$. Furthermore, the derivative of the miss distance will also grow exponentially and at the desired terminal time both will be very large even for a small error in Ko'. This, compounded by limited machine precision, prevents Newton's Method from converging in a reasonable length of time and renders simple shooting unsuitable for most economic problems.

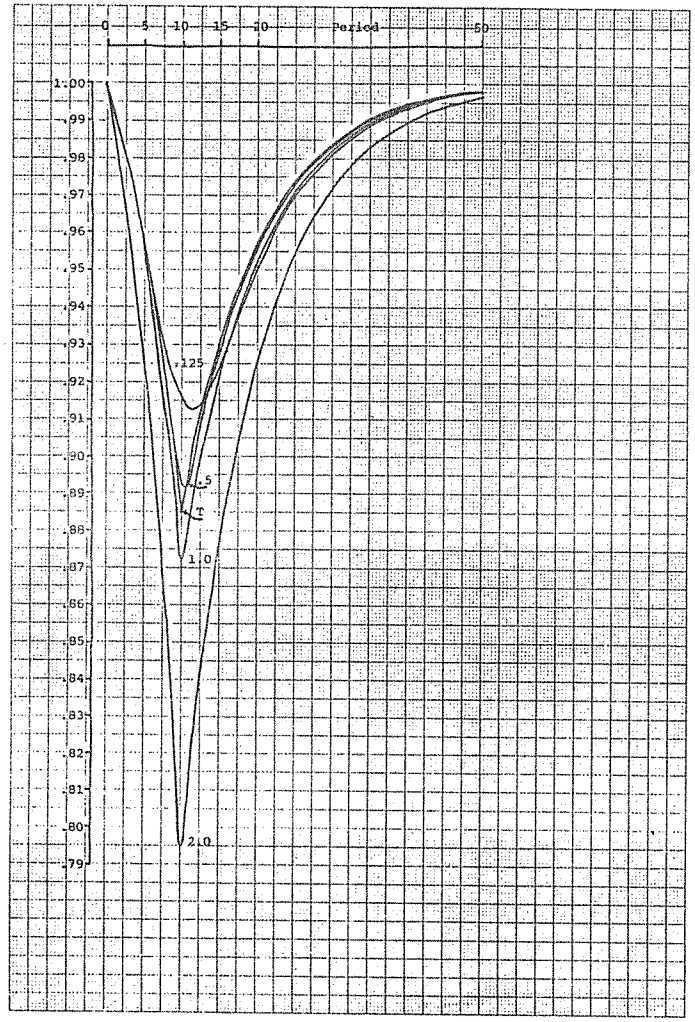
5.4.3 Policies of Short Duration

The final consideration related to grid structure is the duration of the policy to be simulated. To obtain a useful solution, it is necessary to have a fairly large number of points in the period of interest, so brief policies will require a high local point density during the time the policy is in effect.

A related issue is the location of grid points near the onset of a policy. Specifically, it is essential to have a grid point located at the actual time the policy takes effect. Otherwise, if the grid density is increased, the effective time of implementation will shift. For example, if a tax is to commence at time 15 and the nearest points are at 10 and 20, the solution will appear as though the policy began at time 20. If the grid density is later doubled, placing a point at period 15, the effective time of implementation will be 5 periods earlier. The two solutions will not be comparable as they result from fundamentally different policies. This seems a small point, but the consequences of overlooking it can be quite large, as the optimal path is usually fairly sensitive to the time of implementation.

Thus, while the finite difference method is easy to apply, some care must be exercised to generate useful results. As suggested above, the terminal approximation to the steady state is not very important, but grid structure certainly is. Non-uniformity of grid spacing is a particularly powerful method of improving numerical results when some information about the form of the solution is available. Also, discontinuities in the exogenous variables are easily handled if attention is given to approximating their derivatives, and if grid points are located at the appropriate implementation times.

Figure 5.4.1.1: The Effect of ρ in a Logistic Function Dividend Tax



Shown are trajectories of the capital stock under various values of ρ.

3.2 Multiple Shooting

Multiple Shooting is a refinement of simple shooting that helps to control the explosive tendency of numerically unstable models. Basically, the full period of the problem is divided into a number of subintervals and the model is shot over each. This means that rather than searching for a single missing initial condition, the algorithm must find a vector of conditions spread out across the shooting points. Updating the guess vector becomes considerably more complicated, but breaking the problem into subintervals provides a great deal of control over unstable problems.

As a simple example of how multiple shooting is used, consider solving the model above over two adjoining intervals: $[T^0,T^1],[T^1,T^2]$. The capital stock at T^0 is known to be Ko, and its value at T^2 is required to be K^2; the value of Ko´ is unknown. To use multiple shooting, a guess vector of $\tilde{K}o'$, $\tilde{K}^1$ and $\tilde{K}^{1\prime}$ is generated and the model integrated over each subinterval. Denoting the guesses by vector $\tilde{K} = (\tilde{K}o',\tilde{K}^1,\tilde{K}^{1\prime})$, the following set of miss distances can be computed:

$$M^1 = \tilde{K}^1 - \kappa(T^0,T^1,Ko,\tilde{K}o') ,$$

$$M^2 = \tilde{K}^{1\prime} - \kappa'(T^0,T^1,Ko,\tilde{K}o') ,$$

$$M^3 = K^2 - \kappa(T^1,T^2,\tilde{K}^1,\tilde{K}^{1\prime}) .$$

Expanding each about $\tilde{K}$ for a trial solution vector $\hat{K} = (\hat{K}o',\hat{K}^1,\hat{K}^{1\prime})$ gives, in matrix notation, the following system:

$$M(\hat{K}) = M(\tilde{K}) + \left(\frac{\partial M}{\partial K} \right) \cdot (\hat{K}-\tilde{K}) .$$

Since $\hat{K}$ is a solution, $M(\hat{K})=0$ and the equation can be rearranged to obtain the updating formula below:

$$\hat{K} = \tilde{K} - \left(\frac{\partial M(\tilde{K})}{\partial K}\right)^{-1} \cdot M(K) \ .$$

Thus, updating the guess vector now requires solving a system of equations. If the number of variables or the number of shooting points is substantial, the system will be quite large. In sum, multiple shooting consists of the following steps: (1) guess a vector of missing initial conditions, (2) compute (by integration) the consequent miss distances and their partial derivatives, (3) update the guess and try again. This method is described very clearly in Roberts & Shipman (1972), and also in Lipton, et al., (1982). In economic models, shooting is generally much too unstable to be useful; multiple shooting, however, can be used to control the instability simply by increasing the number of shooting points. Accuracy is determined by the tolerance allowed in the miss distances.

Shooting and multiple shooting are sometimes referred to as indirect methods of solving two-point boundary value problems since the terminal time conditions are satisfied (approximately) by finding an initial value problem whose solution at the terminal time is close to the desired value. Direct methods (which do not involve solving an initial value problem) are available and one such technique is finite differences.

3.3 Finite Differences

The theory behind the finite difference method is much simpler than either shooting or multiple shooting: all that is required is to replace all differentials in the model by their finite difference equivalents. This produces a system of equations which can be solved simultaneously to satisfy the boundary conditions exactly. Furthermore, the finite difference method is not subject to the explosive instability that complicates the use of shooting

Overall, discontinuities in the exogenous variables will make both analytic and numerical solution more difficult. The problems are exacerbated by the use of second-order form and this is an important reason why it will often be easier to solve the model as a system of first-order equations. The second-order form can be used, however, if attention is given to selecting a grid compatible with the approximations made for the derivatives of exogenous variables.

The source of this curious result is relationship between grid spacing ahd the value of ρ. Notice that finding the solution will require, in essence, integrating the approximated differentials. When ρ is small, the solution becomes inaccurate because the tax change is spread out over a long period of time. As ρ is increased, the solution improves because the approximation to a true step function is improving. After a point, however, the entire transition of the step occurs between τ-h and τ+h. Further increases in ρ will raise the derivative of the tax at τ and decrease the transition time even more. Integration, however, is limited by the grid spacing, so the integral will behave as though the derivative at τ prevailed for all of $[\tau-h,\tau+h]$. Thus if ρ is increased beyond the point at which the jump takes only 2h, the solution will be degraded. Grid spacing, therefore, places an upper bound on the steepness of continuous approximations to discontinuous functions.

A simpler and more reliable technique for handling derivatives is to replace them with their finite difference equivalents; the dividend tax, for example, becomes:

$$T^{d\prime}(t) \cong \frac{T^d(t+h) - T^d(t-h)}{2h} .$$

This has the desirable property that the sharpness of the change increases with grid density; also, the solution will not be degraded by having an excessive value of g on sparse grids. This approximation must be used with care, however, as it tends to smooth out the solution; the computed derivative of a function which jumps once at τ will be non-zero not only at τ, but also at τ-h since the difference formula uses adjacent points. This is the method used in all simulations previously discussed, so it is clear that if the grid is chosen properly, this approximation is very useful: smoothing of the solution is limited.

techniques, so it is particularly useful for economic problems. Finally, the use of finite differences requires only a single solution to a system of equations which, for many models, will be linear or easily linearized. The main cost of all these advantages is the additional effort required to put the model in finite difference form. This technique will be discussed in detail in the following sections.

4. The Finite Difference Method

The use of finite differences is straightforward and can be divided into the following steps. First, a finite difference approximation is selected for each differential in the model. Second, the model is reformulated by replacing all differentials with their approximations. This produces an equation which represents the original model at a specific point in time. The full model will require a system of such equations, each holding at a different time, so the third step in using finite differences is to select a vector of times (referred to as a "net" or "grid" of points), at which the model will be approximated. Fourth, the boundary conditions are applied by setting particular variables appropriately. Finally, the system of equations, now devoid of differentials, is solved. If the original model was linear, the resulting system of equations will also be linear and the solution can be found easily using any of a number of available techniques. If it was nonlinear, the finite difference system will also be nonlinear and may be solved by Newton's Method, or it may be linearized and solved directly. In the current section linear models will be considered; nonlinearity will be discussed in section 6.

4.1 Difference Formulae

Selecting the difference formulae to be used is of some importance to the success of the method and requires a compromise between ease of model

formulation and computational load. On a uniform grid of points located a distance h apart, a typical first-order central difference formula is shown below:

$$f'(t) \cong \frac{f(t+h) - f(t-h)}{2h} ;$$

its second-order counterpart is:

$$f''(t) \cong \frac{f(t+h) - 2f(t) + f(t-h)}{h^2} .$$

These expressions are truncated Taylor series expansions with error terms $O(h^2)$. It is possible to construct difference formulae with higher order accuracy, but at the cost of considerable additional complexity. Since the truncation error can be made arbitrarily small by using smaller and smaller step sizes h, it is possible to use either a simple difference formula and a dense grid, or a more complex formula and a sparse grid. In general it is probably best to use simple formulae, such as those above, and a small step size, unless computing resources are severly constrained. For more detail on alternative difference formulae, refer to the appendix.

Overall, results obtained with this technique will be somewhat inaccurate because of truncation error arising from dropping higher-order terms in the Taylor series and because of roundoff error introduced by limited computing precision. Using currently available equipment the former source of error will generally dominate and a number of methods by which it can be reduced will be discussed in section 5.

Finally, it should be noted that the finite difference approximation converges to the true solution as the step size approaches zero. This can be proved; one such proof appears in Isaacson and Keller (1966).

5.4.2 Derivatives

Sometimes it will be necessary to approximate the derivative of an exogenous variable at a discontinuity, particularly if the model is to be solved in second-order form (conversion will require differentiating one of the first-order conditions with respect to time). This leads naturally to the question of whether it is possible to eliminate the problem by approximating the discontinuous change with a continuous function. A step in the dividend tax from 0 to z at time τ, for example, can be approximated by the logistic function:

$$T^d(t) = \frac{z}{1+\exp(-4\rho(t-\tau))} .$$

As time goes to $-\infty$, the tax goes to zero; as time goes to $+\infty$, it approaches z; at time τ, the tax is z/2 and its derivative is just ρ. The function approximates a step from 0 to z at time τ, the sharpness of the transition being determined by parameter ρ. With such an approximation it would seem possible to obtain a result arbitrarily close to a discontinuous change simply by increasing ρ. Because of an interaction between grid spacing and the value of ρ, however, this turns out not to be the case and it is quite easy to degrade the numerical result to the point of absurdity by inappropriate choice of ρ for a given grid. This is demonstrated clearly in figure 5.4.1.1 which shows the results of conducting the standard dividend tax experiment using a logistic function to approximate the discontinuous tax change. In each of the simulations shown, z=.25 and τ=10; several values of ρ were used to test the effect of increasing the sharpness of the transition. Clearly, increasing ρ improves the solution only while $\rho \leq .5$; slightly above that further increases degrade it, and even the ρ=1 solution is inferior to that for ρ=.5.

will not preclude finding the problem's first-order conditions. Optimal control requires that the function $f(I,K)=I-\dot{o}K$ and its partial derivatives be continuous functions of I and K, and that the control variable, I, be piecewise continuous. Under these conditions the solution generated will be continuous and piecewise smooth as a function of time (see Leitman (1981), p. 81). Fortunately, the conditions are satisfied even when exogenous variables change discontinuously, so the method of optimal control will remain applicable.

The second question - how discontinuities are handled when solving the first-order conditions explicitly - is more difficult to answer. In practice the effect of a discontinuity in an exogenous variable will be to produce a discontinuity in the control variable and hence a jump in the derivative of the state variable; this is illustrated clearly by the behaviour shown in figure 5.1.1 above. The results displayed were obtained analytically by a method discussed in the descriptive paper; here attention will be focused on the implications of discontinuities for numerical analysis.

A model with a discontinuous exogenous variable is numerically equivalent to a continuous problem with a region of extremely high curvature near the time of implementation. From the discussion of grid density above, it is clear that truncation error will be large at implementation unless a grid of fairly high density is used. The error will be manifest as smoothing of the solution near the discontinuity, and can be reduced easily by increasing the local grid density. A related, and more serious, problem arises when the derivative of a discontinuous function must be evaluated, as discussed in the following section.

4.2 Finite Difference Form of a Second-Order Model

As an example of how the method is used, the following problem will be discussed: find function $f(t)$, $t\varepsilon[T_a,T_b]$, such that

$$a(t)f''(t) + b(t)f'(t) + c(t)f(t) = d(t) , \qquad (4.2.1)$$

and satisfying boundary conditions:

$$f(T_a) = f_a \qquad \text{and} \qquad f(T_b) = f_b .$$

Inserting central difference approximations into equation (4.2.1) gives its finite difference equivalent:

$$a(t)\left(\frac{f(t+h)-2f(t)+f(t-h)}{h^2}\right) + b(t)\left(\frac{f(t+h)-f(t-h)}{2h}\right) + c(t)f(t) = d(t) .$$

Finally, collecting terms in f results in:

$$f(t-h)\left(\frac{a(t)}{h^2} - \frac{b(t)}{2h}\right) + f(t)\left(c(t) - \frac{2a(t)}{h^2}\right) + f(t+h)\left(\frac{a(t)}{h^2} + \frac{b(t)}{2h}\right) = d(t) ,$$

where h is related to the number of grid intervals, n, by:

$$h = \frac{T_b-T_a}{n+1} .$$

The finite difference form above describes the relationship between f(t) and the adjacent points of the function, so to find a complete solution requires solving a system of such equations simultaneously. It is convenient to introduce a variable t^i defined by:

$$t^i = T_a + i\cdot h, \qquad i=0,\ldots,n+1 ,$$

and to define the functions α, β and γ as shown:

$$\alpha(t) = \left(\frac{a(t)}{h^2} - \frac{b(t)}{2h}\right), \quad \beta(t) = \left(c(t) - \frac{2a(t)}{h^2}\right), \quad \gamma(t) = \left(\frac{a(t)}{h^2} + \frac{b(t)}{2h}\right) .$$

The full model may now be written as a system of equations:

$$\begin{bmatrix} \alpha(t^1) & \beta(t^1) & \gamma(t^1) & 0 & & \\ 0 & \alpha(t^2) & \beta(t^2) & \gamma(t^3) & 0 & \\ & & \vdots & & \vdots & \\ & & 0 & \alpha(t^n) & \beta(t^n) & \gamma(t^n) \end{bmatrix} \begin{bmatrix} f(t^0) \\ f(t^1) \\ \vdots \\ f(t^{n-1}) \\ f(t^n) \\ f(t^{n+1}) \end{bmatrix} = \begin{bmatrix} d(t^1) \\ d(t^2) \\ \vdots \\ d(t^n) \end{bmatrix} .$$

The boundary conditions require that $f(t^0) = f_a$ and $f(t^{n+1}) = f_b$. These can be applied by partitioning the left side and rearranging to get:

$$\begin{bmatrix} \beta(t^1) & \gamma(t^1) & 0 & 0 \\ \alpha(t^2) & \beta(t^2) & \gamma(t^2) & 0 \\ & \vdots & & \vdots \\ & 0 & \alpha(t^n) & \beta(t^n) \end{bmatrix} \begin{bmatrix} f(t^1) \\ f(t^2) \\ \vdots \\ f(t^n) \end{bmatrix} = \begin{bmatrix} d(t^1) \\ d(t^2) \\ \vdots \\ d(t^n) \end{bmatrix} - \begin{bmatrix} \alpha(t^1) & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & \gamma(t^n) \end{bmatrix} \begin{bmatrix} f_a \\ f_b \end{bmatrix} .$$

Finally, this may be written compactly in matrix notation as:

$$\Theta F = D - B ,$$

where $B = (\alpha(t^1)f_a, \ \ldots \ 0 \ \ldots, \ \gamma(t^n)f_b)^T$; Θ, F and D are clear from context. This direct use of the boundary conditions eliminates the need for an iterative search for the equivalent initial value problem and thus avoids the numerical instability associated with shooting methods.

Solving the model requires finding the missing values of f by computing

$$F = \Theta^{-1} \cdot (D-B) .$$

This can be accomplished easily using any of the available methods for solving systems of linear equations. Furthermore, it is a general structure which applies to any second-order linear differential equation for which a(t) is

failing is to perform a preliminary experiment using a uniform grid of moderate density; the results obtained are then used to select a more sparse grid with points in appropriate locations. Overall, non-uniformity in grid point location is a powerful method by which a priori knowledge of the form of the solution can be used to improve the results of numerical analysis. This, and the discussion of the number of points in the previous section, serves to emphasize the importance of grid structure to the accuracy of the numerical solution.

5.4 Exogenous Variables

The quality of the solution also depends heavily on the path of the exogenous variables. Three aspects that must be dealt with carefully are: (1) discontinuities, (2) derivatives in the coefficient functions, and (3) policies of short duration. All of these are closely related to grid spacing and the problems discussed below become unimportant as the grid becomes very dense. Since a fairly sparse grid will often have to be used, these factors are worth discussing in detail.

5.4.1 Discontinuities

In most experiments the policy of interest will involve a change which is not differentiable and may even be discontinuous: in the dividend tax experiment, for example, the tax changes discontinuously at implementation. This brings up two important questions: first, can the optimal control problem still be solved, and second, can the resulting differential equations be solved explicitly for the optimal path?

Addressing the first question, discontinuities in the exogenous variables

For this reason, manipulation of the grid is vital to the feasibility of numerical analysis because it allows accurate solutions to be obtained using a limited grid. (In the finite difference form of a model, an equation is associated with each grid point, so the computational load of the model increases with the square of the number of points.) Reducing the number of points is particularly important when the foresight model is to be integrated into a larger system that is already close to the computational limits of available equipment. Overall, using a non-uniform grid allows high accuracy to be obtained with few points.

There are, however, at least two difficulties with the use of a non-uniform grid. First, most mathematical work on convergence and error bounds for numerical analysis is based on uniformly spaced grid points; it may be impossible to derive a realistic a priori error bound for a non-uniform grid. This is not a problem, however, as long as some method is available to estimate truncation error. In the uniform grid case, this can be done by generating several simulations with increasing grid density; comparing the results will indicate the rate of change of the solution with respect to step size. Increasing the density of a non-uniform grid is less easy to define, but simply placing an additional point between every existing point is one reasonable approach that will give some idea of the size of the truncation error. Unfortunately, it is not clear that any sort of extrapolation can be used to refine such results.

The second and more important problem with use of a non-uniform grid is the danger of concentrating grid points away from regions of high curvature: this will degrade the solution for the same reason that the opposite movement improves it. One approach to grid selection that will not be subject to this

never zero. (If at some time ζ, $a(\zeta) = 0$, the equation is said to possess a singular point at ζ and the analysis presented here would have to be modified.) Assuming there are no singularities on the domain of interest, the general structure above may be used for any particular differential equation by inserting the appropriate functions a, b, c and d. This makes it possible to use a single solution routine with a variety of related models.

4.2.1 Example: The Constant Returns to Scale Model

To find the finite difference form of the constant returns model, all that is necessary is to substitute the appropriate functions into the general second-order form given above. Recall that the second-order form of the CRTS model, as derived in section 2, is:

$$K'' + K'(g-r) + \delta K(g-r-\delta) = \frac{1}{2\theta}\left((1+\gamma)(r+\delta-g) - \frac{\beta}{P_k(1-T^s)}\right) .$$

Comparison of this formula with the general form given in equation (4.2.1) shows that:

$$a(t) = 1 ,$$

$$b(t) = g-r ,$$

$$c(t) = \delta(g-r-\delta) ,$$

$$d(t) = \frac{1}{2\theta}\left((1+\gamma)(r+\delta-g) - \frac{\beta}{P_k(1-T^s)}\right) .$$

Substituting these into the general finite difference form gives the following:

$$K_{t-h}\left(\frac{1}{h^2} - \frac{g-r}{2h}\right) + K_t\left(\delta(g-r-\delta) - \frac{2}{h^2}\right) + K_{t+h}\left(\frac{1}{h^2} + \frac{g-r}{2h}\right) = \frac{1}{2\theta}\left((1+\gamma)(r+\delta-g) - \frac{\beta}{P_k(1-T^s)}\right) .$$

All that is required to put a model in finite difference form is to replace the

differentials with finite difference approximations, but it is convenient to simplify the resulting equation by collecting terms in the state variable (in this case K), as was done above.

4.3 Invertability

One important question is whether the inverse of matrix Θ exists. The band-diagonal structure of Θ makes this likely as long as the functions α, β, and γ are reasonably well-behaved, and in most problems no difficulty will arise. A sufficient condition, however, for the existence of an inverse is that Θ be strictly diagonal dominant, which is true when the following holds (where abs(•) is the absolute value function):

$$\text{abs}(\,\Theta_{i,i}\,) \;>\; \sum_{j\neq i}^{n} \text{abs}(\,\Theta_{i,j}\,), \qquad i = 1,\ldots,n\ .$$

In the current context this requires that:

$$\text{abs}\Big(\,c(t) - \frac{2a(t)}{h^2}\,\Big) \;>\; \text{abs}\Big(\,\frac{a(t)}{h^2} - \frac{b(t)}{2h}\,\Big) \;+\; \text{abs}\Big(\,\frac{a(t)}{h^2} + \frac{b(t)}{2h}\,\Big)\ ,$$

hold for all t. Since there are no singularities in the domain of t, it may be assumed without loss of generality that the equation has been normalized so that $a(t)\equiv 1$. If, in addition, $c(t) < 0$ then the existence of an inverse is guaranteed when $h < 2/\text{abs}(b(t))$ for all $t\varepsilon[T_a,T_b]$. This is a sufficient condition only; in practice an inverse will exist in a much broader range of circumstances. The above condition does provide a simple test, however, and also illustrates the importance of the grid size h.

The results are striking: using only nine grid points at carefully chosen locations, it is possible to obtain a solution as accurate as that generated by a uniform grid of 159 points (with the exception of the value at time 10 which is slightly less accurate). The error at all points except 10 is negligible, and the error at 10 has been reduced from 6.1% to only .8%. Furthermore, the drop in K at implementation has been brought to 94% of its true value.

To see why this is so effective, recall that each finite difference approximation was constructed from several Taylor series expansions such as that shown below:

$$f(a+\varepsilon) = f(a) + f'(a)\varepsilon + \frac{f''(a)\varepsilon^2}{2!} + \frac{f'''(a)\varepsilon^3}{3!} + \ \ldots\ .$$

Ignoring terms above second order, truncation error is greatest where f'' is large, so an upper bound on the error is:

$$\varepsilon^2 \cdot \sup\{\, f''(t),\ t\varepsilon[T_a,T_b]\,\}/2\ .$$

A non-uniform grid decreases truncation error by increasing ε where f'' is small and decreasing it where f'' is large; roughly, it would be ideal to choose $\varepsilon(t)$ to solve:

$$\min \sup\{\, \varepsilon^2(t)\cdot f''(t),\ t\varepsilon[T_a,T_b]\,\}\ ,$$

but, since f'' is unknown, this is impossible. As demonstrated by the results above, however, any movement of grid points from regions of low curvature to regions of high curvature is a practical measure which will improve the solution substantially. (Although this example illustrates why nonuniformity is a powerful tool, in the difference formulae acutally used it is f''' rather than f'' that is relevant for truncation error; for further discussion, refer to the appendix.)

to 60, with the effects discussed above, or (2) a non-uniform grid could be used which had many points before 60 and few after it. Such non-uniform grids are the subject of the next section.

5.3.2 Grid Spacing

Non-uniformity of grid spacing is an extremely powerful tool for improving the accuracy of numerical solutions. All that is required is some a priori knowledge of where the curvature of the solution is likely to be largest; extra points are located there and removed from regions of low curvature. The effect of this can be demonstrated by varying the spacing of the n=9 grid discussed above. The results of the dividend tax experiment are displayed in table 5.3.2.1 for several non-uniform grids; the terminal condition was imposed at time 100 in all cases. Each grid had two terminal and nine interior points located at the times for which values are shown.

Table 5.3.2.1: The Effect of Grid Spacing

Time	True Value	Grid a	b	c	d	e
0.0	1.000	1.	1.	1.	1.	1.
5.0	.958	--	.980	.991	.999	1.001
7.0	.934	--	--	.985	.999	1.001
9.0	.904	--	--	--	.997	1.000
9.5	.895	--	--	--	--	1.000
10.0	.885	1.061	1.032	1.020	1.010	1.008
20.0	.958	1.019	1.008	1.004	1.001	1.
30.0	.984	1.007	1.003	1.001	1.	1.
40.0	.994	1.002	1.001	1.	1.	1.
50.0	.998	1.001	1.	1.	1.	1.001
60.0	.999	1.	1.	1.	1.	--
70.0	1.000	1.	1.	1.	--	--
80.0	1.000	1.	1.	--	--	--
90.0	1.000	1.	--	--	--	--
100.0	1.000	1.	1.	1.	1.	1.

(Figures shown are the ratio of the numerical result to the true value.)

4.4 Systems of Equations

The above discussion was conducted in terms of a model specified by a single second-order differential equation, but the method can be applied much more widely. In particular, it is straightforward to solve systems of several equations, so the first-order conditions need not be converted to second-order form in order to be solved. The second-order form does, however, provide a certain amount of intuition which the first-order system does not, and the discussion of finite difference implementation is somewhat clearer for a single equation. It is useful now to set up the same problem in first-order form as an example.

Recall the model discussed in section 2, which had the first-order equations below:

$$\lambda = P_k(1+\gamma+2\theta I)(1-T^d)(1-T^s) ,$$

$$\lambda' = (r+\delta)\lambda - \beta(1-T^d) ,$$

$$K' = I - \delta K .$$

Using the third equation to eliminate investment gives the following:

$$\lambda' = (r+\delta)\lambda - \beta(1-T^d) ,$$

$$K' = \left(\frac{\lambda}{2\theta P_k(1-T^d)(1-T^s)}\right) - \delta K - \left(\frac{1+\gamma}{2\theta}\right) .$$

Defining appropriate auxiliary functions allows the system to be written:

$$\lambda'(t) = a(t)\lambda(t) + b(t)K(t) - c(t) ,$$

$$K'(t) = d(t)\lambda(t) + e(t)K(t) - f(t) .$$

(Note that in the model above, $b(t)\equiv 0$ and functions a and e are independent of

time.) Putting this in finite difference form gives:

$$\frac{\lambda(t+h) - \lambda(t-h)}{2h} = a(t)\lambda(t) + b(t)K(t) - c(t) ,$$

$$\frac{K(t+h) - K(t-h)}{2h} = d(t)\lambda(t) + e(t)K(t) - f(t) .$$

Finally, this can be written as:

$$\begin{bmatrix} -1/2h & 0 & -a(t) & -b(t) & 1/2h & 0 \\ 0 & -1/2h & -d(t) & -e(t) & 0 & 1/2h \end{bmatrix} \begin{bmatrix} \lambda(t-h) \\ K(t-h) \\ \lambda(t) \\ K(t) \\ \lambda(t+h) \\ K(t+h) \end{bmatrix} = \begin{bmatrix} -c(t) \\ -f(t) \end{bmatrix} .$$

This system expresses the relationship between $(\lambda(t),K(t))$ and the adjacent points, so a complete solution requires a set of such systems, one for each grid point. This final set of equations has the structure shown below:

$$\begin{bmatrix} -1/2h & 0 & -a(t^1) & -b(t^1) & 1/2h & 0 & & \\ 0 & -1/2h & -d(t^1) & -e(t^1) & 0 & 1/2h & & \\ & & -1/2h & 0 & -a(t^2) & -b(t^2) & 1/2h & 0 \\ & & 0 & -1/2h & -d(t^2) & -e(t^2) & 0 & 1/2h \\ & & & & \vdots & & & \\ & & & & & & & \end{bmatrix} \begin{bmatrix} \lambda(t^0) \\ K(t^0) \\ \lambda(t^1) \\ K(t^1) \\ \vdots \\ \lambda(t^{n-1}) \\ K(t^{n-1}) \\ \lambda(t^n) \\ K(t^n) \\ \lambda(t^{n+1}) \\ K(t^{n+1}) \end{bmatrix} = \begin{bmatrix} -c(t^1) \\ -d(t^1) \\ -c(t^2) \\ -c(t^2) \\ \vdots \\ -c(t^n) \\ -d(t^n) \end{bmatrix} .$$

Of the vector of λ and K variables, the values of two will usually be known from the boundary conditions, so the system can be partitioned in a manner

Table 5.3.1.2: Convergence of Solution with Increasing Grid Density

	Grid Density (n)					
	9	19	39	79	159	True
K(10)	.939	.909	.895	.889	.886	.885
Change		.030	.014	.006	.003	

This rate of decrease is important for assessing the accuracy of a numerical solution when true values cannot be computed: the magnitude of the truncation error can be estimated using only numerical results. Furthermore, it is possible to use Richardson's Extrapolation (see Birkhoff and Rota (1969), p. 214) to exploit the convergence of numerical values resulting from increased grid density. All that is required is to compute solutions for several grid densities; these can be combined to yield a result more accurate than any of the components.

In each simulation the largest inaccuracy occurs at the time of implementation. This occurs because at that point the dividend tax changes extremely rapidly (actually, it is discontinuous), and the extent to which the numerical solution can accurately represent sudden changes is limited by the spacing of the grid points, a phenomenon which is discussed in more detail in the next section.

One final observation on the results in table 5.3.1.1 is that all of the simulations reach the steady state (to three decimal places) by period 60. This suggests that little is achieved by putting 40% of the grid points in the region from 60 to 100. Computing costs depend solely on the number of points used and not on their location in time, so either of the following measures could be taken to improve efficiency: (1) the terminal time could be moved up

which the simulations were run, so truncation error will be dominant. The largest error occurs in the period 10 results; for the n=9 simulation it is 6.1%. If the model is intended to estimate the size of the capital stock, this error is fairly small; on the other hand, if its purpose is to estimate the deviation of the capital stock from its initial level, the percentage error is much larger. In this experiment the actual drop in K(10) is .115, while the numerical simulation returns a drop of .061, so the magnitude of numerical drop is only 53% of the true value.

Increasing the density to n=19 divides the error in K roughly in half, to 2.7%, and brings the drop to 79% of its true size, so whichever way accuracy is measured, doubling the number of points cuts the approximation error by about half, as suggested above. Doubling grid density again moves the numerical drop to 92% of its true value; the error in K is reduced to 1.1%. In these experiments the main effect of economic interest is the drop in the capital stock, rather than its level, so the appropriate test of accuracy is the model's ability to reproduce this decline over the pre-implementation period. Under this criterion, the n=9 solution is not adequately accurate, as it underestimates the drop in the capital stock by 47%. In contrast, the n=39 simulation is only 8% low by this standard, and it would be satisfactory if computing limitations prevented the use of additional grid points.

Finally, notice that the difference between numerical values for a particular time is decreasing as n increases. The numerical results at period 10 are given below for a number of different grid densities; the change in the value of K as n is increased is also shown.

analogous to that used in section 4.2. Typically the boundary values will have the form below:

$$K(t^0) = K(T_a) = K_a ,$$
$$K(t^{n+1}) = K(T_b) = K_b ,$$

and the corresponding values of $\lambda(t^0)$ and $\lambda(t^{n+1})$ will be found as part of the solution. (This means that two extra equations will be required to make the Θ matrix square: a reverse difference equation for $\lambda(t^0)$ in terms of $\lambda(t^1)$ and $\lambda(t^2)$, and a similar equation for $\lambda(t^{n+1})$ in terms of its predecessors. For more detail, consult Isaacson and Keller.)

Using a system of equations is particularly useful when the order of the model is greater than two. Formally, any n^{th}-order nonlinear differential equation may be converted to a system of n nonlinear first-order differential equations by a simple transformation, and this provides a valuable way of reducing the complexity of the finite difference form of a model. The transformation is accomplished as follows. Consider the arbitrary n^{th}-order equation:

$$f^n(t) = g(t,f,f',f'',f''',\ldots,f^{n-1}) .$$

New functions are defined as shown:

$$f_1 = f ,$$
$$f_2 = f_1' = f' ,$$
$$\vdots$$
$$f_n = f_{n-1}' = f^{n-1} .$$

Substituting these into the original equation and rearranging gives a system of n first-order equations:

$$f_1' = f_2 ,$$
$$f_2' = f_3 ,$$
$$\vdots$$
$$f_{n-1}' = f_n ,$$
$$f_n' = g(t,f_1,f_2,f_3,\ldots,f_n) .$$

This has the advantage of limiting the difference formulae required to only that for first-order, but it also doubles the number of equations which must be solved for a grid of given density. All models discussed in this paper were computed using the second-order form, but it should be emphasized that the finite difference technique does not restrict the form of the model.

5. Practical Considerations

To implement the method successfully it is important to understand the factors which improve or degrade the numerical result relative to the true solution. Several of these are: (1) specification of boundary conditions, (2) grid structure, and (3) coefficient functions. Each of these will be discussed in detail, and sample numerical results will be presented for the model described in the following section.

5.1 Constant Returns Model: The Dividend Tax Experiment

The model used will be that described in section 2, the second-order form of which is repeated below:

$$K'' + K'(g-r) + \delta K(g-r-\delta) = \frac{1}{2\theta}\left((1+\gamma)(r+\delta-g) - \frac{\beta}{P_k(1-T^s)}\right) ,$$

the true value of the capital stock and then the ratio of each numerical solution to it. Recall that the experiment was an announced increase in the dividend tax from 0 to 25% to take effect in period 10.

Table 5.3.1.1: The Effect of the Number of Grid Points

Time	True Value	Number of Grid Points 9	19	39	79	159
0	1.000	1.000	1.000	1.000	1.000	1.000
5	.958	-	.979	.991	.995	.997
10	.885	1.061	1.027	1.011	1.005	1.001
15	.930	-	1.015	1.006	1.003	1.001
20	.958	1.019	1.008	1.003	1.001	1.000
25	.974	-	1.005	1.003	1.001	1.001
30	.984	1.007	1.003	1.002	1.001	1.001
35	.991	-	1.001	1.000	1.	1.
40	.994	1.005	1.001	1.001	1.	1.
45	.997	-	1.	1.	1.	1.
50	.998	1.001	1.	1.	1.	1.
55	.999	-	1.	1.	1.	1.
60	.999	1.	1.	1.	1.	1.
65	1.000	-	1.	1.	1.	1.
70	1.000	1.	1.	1.	1.	1.
75	1.000	-	1.	1.	1.	1.
80	1.000	1.	1.	1.	1.	1.
85	1.000	-	1.	1.	1.	1.
90	1.000	1.	1.	1.	1.	1.
95	1.000	-	1.	1.	1.	1.
100	1.000	1.	1.	1.	1.	1.

(Figures shown are the ratio of the numerical result to the true value.)

Several points are evident from these results: (1) the error is fairly small, even for a sparse grid, (2) the error declines rapidly as grid density is increased, (3) the largest error occurs at implementation, and (4) nothing is gained by continuing the solution past period 60. Each of these will be dealt with briefly below.

The total error in a numerical result is composed of truncation error and roundoff error. The latter is less than 1 part in 10^6 for the computer on

5.3 Grid Structure

Accuracy of the numerical solution depends heavily on the number and spacing of grid points: the greater the density of points, the more accurate the solution. Unfortunately, size of the system of equations goes up with the square of the number of points and the number of steps required for a solution rises even more rapidly, so attainable accuracy will usually be constrained by computational limitations. For this reason, the structure of the grid should be chosen carefully for maximum accuracy at minimum cost. The aspects to be considered are: (1) the actual number of points to be used, and (2) their locations in the interval $[T_a,T_b]$.

5.3.1 Number of Grid Points

As noted in section 4.2, when grid points are uniformly distributed over the interval the number of points is related to the spacing between them by:

$$h = \frac{T_b - T_a}{n+1} .$$

Because the finite difference formulae have truncation error $O(h^2)$, cutting the step size in half reduces the error considerably. The error made at each step is $O(h^2)$ and the number of points is proportional to $1/h$, so the cumulative error at any given step is $O(h)$. Thus, halving the step size should result in halving the truncation error (see Birkhoff and Rota (1969), p. 188).

To illustrate the importance of the number of grid points the dividend tax experiment was solved for n=9, 19, 39, 79, and 159. The time interval used was [0,100] in all cases, so the step sizes were h=10, 5, 2.5, 1.25 and .625, respectively. Results are presented in table 5.3.1.1; the figures given are

where

$$g = -\frac{T^{s\prime}}{1-T^s} - \frac{T^{d\prime}}{1-T^d} .$$

Attention will be focused on a single representative experiment: an announced change in the dividend tax rate. The initial tax and the investment subsidy will be zero, the price of capital will be 1.0, and the parameters of the model will have the following values:

$$r = .05, \quad \delta = .10, \quad \theta = 20/3, \quad \gamma = -2/3, \quad \beta = .25 .$$

In the absence of a tax shock the steady state capital stock can be found by evaluating the second-order equation with $K'' = K' = g = 0$; computing shows that the steady state is 1.0 for the parameters above. This implies that the following must be true:

$$\text{Investment} = \delta K^{ss} = .1 ,$$

$$\text{Earnings} = \beta K^{ss} = .25 ,$$

$$\text{Investment Cost} = (1+\gamma)I + \theta I^2 = .1 ,$$

$$\text{Dividends} = E - C(I) = .25 ,$$

$$\text{Market Value} = .15/.05 = 3.0 .$$

The solution will usually be required to attain the steady state at time 100 and the initial capital stock will be 1.0. In the absence of a shock, the model should remain at the steady state, and this is, in fact, the behaviour of the numerical solution.

The dividend tax experiment has the nice feature that its solution can be determined analytically, which is not true of most other experiments. As derived in the descriptive paper, the path of the capital stock after the announcement is given by:

$$K(t) = \frac{\beta(T_1^d - T_2^d)(\exp(-(r+2\delta)(T-t)) - \exp(-(r+2\delta)T)}{2\theta P_k(r+\delta)(r+2\delta)(1-T_1^d)(1-T_1^s)} + K_1^{ss} , \quad (t<T)$$

or

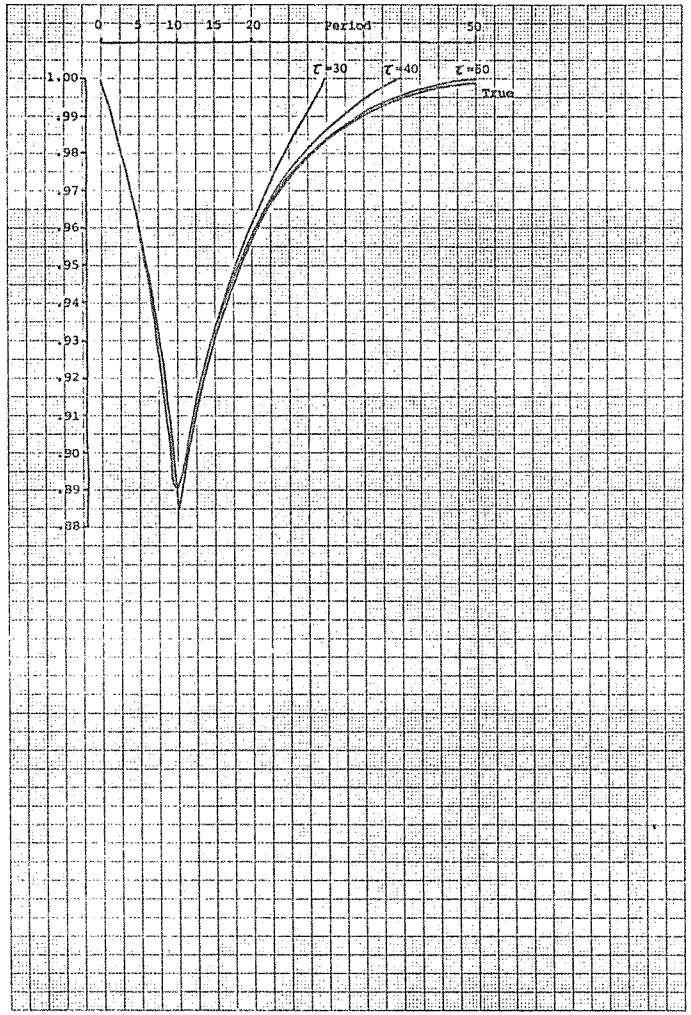
$$K(t) = \left(\frac{\beta(T_1^d - T_2^d)(1-\exp(-(r+2\delta)T)}{2\theta P_k(r+\delta)(r+2\delta)(1-T_1^d)(1-T_1^s)}\right)\exp(-\delta(t-T)) + K_2^{ss} , \quad (t>T)$$

where T is the time at which the tax is to be implemented; T_1^d and T_2^d are its values before and after the change; and K_1^{ss} and K_2^{ss} are the steady states of the model when taxes T_1^d and T_2^d, respectively, are expected to hold from the planning date forever (in this model, both steady states are 1.0).

The solution to the experiment of increasing the tax to 25% at time 10 is plotted in figure 5.1.1. The behaviour is discussed in detail in the descriptive paper; briefly, it is optimal for the firm to reduce investment and pay high dividends just before the tax is implemented.

A number of important features of the finite difference method will be discussed in the following sections, and numerical results will be presented for this experiment to illustrate the accuracy of the approximation under various choices of grid structure and other variables. It is worth noting that the coefficient functions are discontinuous at T, so neither K′ nor K′′ exist there. This presents a formidable challenge for numerical analysis because such techniques will, in essence, generate solutions which are twice continuously differentiable. The effect of this will be to smooth out discontinuities in K′ and hence round out corners in K, such as that at T, so a technique which produces an accurate solution to this experiment is fairly powerful. The finite difference method is capable of such accuracy, provided it is applied carefully with attention to the details discussed below.

Figure 5.2.1: The Effect of Terminal Time on the Solution

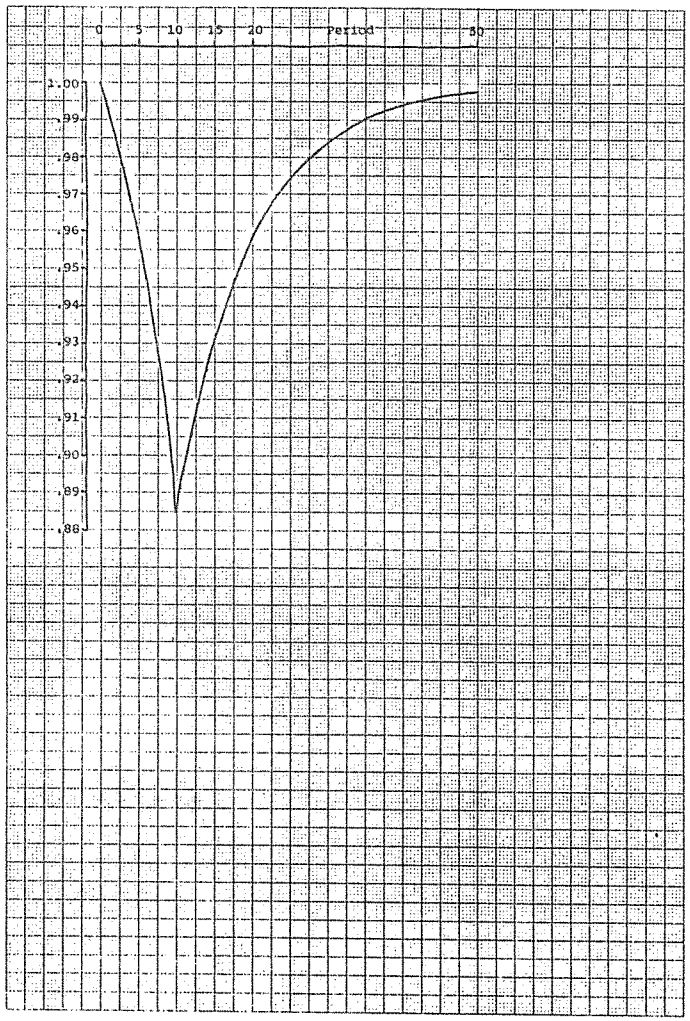


Shown are trajectories of the capital stock under the indicated terminal time conditions.

the true solution except near the terminal time. The latter property means that a reasonably small error in the terminal condition will make very little difference to the result in early periods. The intuition behind this is clear and compelling: because of the positive root in the solution, the only way to attain a point near equilibrium at large values of t is to follow a path close to the optimum for most of the interval. This also illustrates why shooting methods are so difficult to apply: to four digits the initial derivative of the capital stock in the τ=30 and τ=50 solutions is the same, but a difference of 1.5% occurs in the value of K at τ; clearly solutions with initial derivatives which differ by even a small amount will be extremely far apart just 30 periods in the future. (Recall that shooting is essentially an iterative procedure for finding the correct value of K´(0) using the miss distance at large times.) Thus, when the terminal condition is imposed reasonably close to its true value, and at a large time τ, the solution for early periods will be essentially independent of the error in the terminal condition. As a practical matter, if the derivative of the state variable is reasonably close to zero at the terminal point, approximation error generated by the boundary conditions will be unimportant.

This discussion has focused on the divergence of different numerical solutions from each other, but all three of those considered above differ slightly from the true solution. This inaccuracy, as discussed in the following section, is much more serious than terminal time error and arises from the structure of the grid of points used.

Figure 5.1.1: The Dividend Tax Experiment



Shown is the trajectory of the capital stock for an announced increase in the dividend tax from 0 to 25% at period 10. Capital is measured as a fraction of its initial value.

5.2 Terminal Time

Typically economic problems will have some conditions specified at the initial time and others imposed later. If the conditions are given for specific finite times, this presents no particular difficulty: for example, if the solution is required to satisfy:

$$f(T_a) = f_a \qquad \text{and} \qquad f(T_b) = f_b,$$

implementing the conditions is straightforward. Unfortunately, in many problems of interest the conditions given are:

$$f(T_a) = f_a \qquad \text{and} \qquad \lim_{t\to\infty} f(t) = f^{ss} .$$

Because it is impossible to model an infinite length of time, it is necessary to approximate the second condition with one which holds at some large (but finite) time τ. This will lead to an error in the numerical result at τ equal to $f(\tau)-f_\tau$, where f_τ was imposed in lieu of the true value, $f(\tau)$. Furthermore, the error will propagate backward, disturbing the solution slightly at all previous times.

The most expedient way of imposing this constraint is to require the solution to attain the true steady state at some large time τ. In most cases the steady state will be easy to calculate from comparative static analysis, so the main difficulty is selecting τ. To illustrate this, the dividend tax experiment was solved specifying that the true steady state be attained at τ = 30, 40, and 50. The results are plotted in figure 5.2.1, along with the true solution, and numerical results for several points are listed in table 5.2.1.

Table 5.2.1: The Effect of Terminal Time on the Solution

Time	True Value	Capital Stock τ=30	τ=40	τ=50
0	1.0000	1.0000	1.0000	1.0000
2	.9852	.9839	.9839	.9839
4	.9679	.9651	.9650	.9650
6	.9647	.9422	.9420	.9419
8	.9199	.9131	.9128	.9128
10	.8853	.8886	.8881	.8881
12	.9061	.9092	.9084	.9084
14	.9231	.9262	.9251	.9250
16	.9370	.9403	.9387	.9386
18	.9484	.9521	.9499	.9497
20	.9578	.9621	.9591	.9588
22	.9654	.9708	.9666	.9663
24	.9717	.9785	.9729	.9724
26	.9768	.9857	.9781	.9774
28	.9810	.9927	.9824	.9815
30	.9845	1.0000	.9861	.9849

This example illustrates several important properties of the terminal time specification. First, the error at τ appears to be an upper bound on the error at any earlier time. Secondly, the effect of the terminal time on the results for early periods seems to be quite small. Finally, the value of K' at τ provides some indication of the error at that point, and hence the error in earlier periods. The terminal condition problem may also be thought of as choosing the proper value of K to impose at a given time; here it will be convenient to compare solutions which have various values of K imposed at τ=30 (that is, the τ = 50 solution is identical to requiring that K = .9849 at τ=30).

The following conjectures are made but no formal proof has been attempted: (1) the error in the numerical solution at any particular time should be a monotonic function of the error in the terminal specification, and (2) a generalized turnpike property will cause the numerical path to remain close to