Learning and the Stock Flow Model[1]

David K. Levine

Department of Economics

UCLA

April 16, 1996

prepared for Trento Conference in honor of Robert Clower

Abstract: We consider an application of the stock flow model of intertemporal preference to the problem of strategic "learning about learning" with the best response dynamic. We show that when players are myopic, but not as in the usual analysis completely so, that we can find an approximate solution that is "implementable" in the sense that it does not require solving an equilibrium system. Paradoxically, in the steady state, this increase in patience and sophistication by the players generally lowers their utility.

[1] As always I am grateful to conversations with Drew Fudenberg. Peter Howitt read the manuscript carefully and made many useful suggestions. This research was supported in part by NSF grant SBR-9320695.

1. Introduction

In 1954, Robert W. Clower, then an assistant professor at the State College of Washington, Pullman, published a paper in the *American Economic Review* on the dynamics of investment. This article modeled demand for the stock of capital and the flow of investment as jointly depending on the price of capital. From the more current viewpoint this "stock-flow" dynamic was based on a utility function that depends both on the stock of capital and the flow of investment. Current orthodoxy largely rejects this point of view, emphasizing the dependence of utility on all future stocks (or at least the corresponding consumption flows) and prices. However, the actual dynamics of a modern dynamical model can often be reduced to a simple difference equation in the current and next period capital stock, so that the basic features of the investment dynamics discussed by Clower in 1954 are still largely viewed as valid.

There is, however, another viewpoint on why utility might depend both on a stock and a flow: that is because in a world of boundedly rational agents, the flow may be a good proxy for future levels of the stock. (This seems to be the idea in the original work on stock-flow analysis.) In the research reported here, we take that idea seriously to attack one of the most difficult issues in the theory of learning in games: the problem of players learning about each other learning.

In most research on learning (or evolution) in games, it is explicitly or implicitly assumed that players are myopic. Often this is justified by assuming that players meet to play in a large population so that they have little influence on the future. With this assumption the impact of a player on his opponents' learning does not matter because it effects only their future play, about which he does not care. However, in strategic interactions such as repeated play, myopia is not a sensible assumption. Once we drop

that assumption, however, players can have an impact on the future play of their opponents, and potentially they can learn about this at the same time their opponents are learning about them.

To attack the issue of learning about learning, it is necessary to model players who are not myopic, that is who care about, have an impact on, and can learn about their impact on the future. As a simple model of bounded rationality, we reintroduce the stock flow model, assuming that players care not only about the level of utility (as in the myopic model) but also on the rate of change of utility. We then see how the problem of one player learning about another player learning about that player learning leads to an infinite recursion that can be solved only by an equilibrium theory very much against the spirit of the theory of learning. This dilemma of needing equilibrium theory to solve a disequilibrium model is one of our major concerns.

We do not propose a general solution to the dilemma of needing equilibrium theory to solve a disequilibrium model. However, in the special case of small departures from myopia, there is a useful theory; that is, the case in which players are somewhat patient but not terribly patient leads to a simple and sensible theory. In this case the learning about learning about learning is of second order importance, and we can neglect it. Roughly, players learn how they effect the level of utility and how they effect their opponents learning about the level of utility, but not how they effect their opponents learning about their learning about the level of utility, because this is of much less significance. In the case of a small departure from myopia, in other words, we may usefully introduce an approximation by cutting off the recursion after a finite number of steps.

What happens when we cut off the recursion after a finite number of steps? The resulting description of the dynamic interaction between players is a simple variation on the best-response dynamic. The stability theory is much the same as in the myopic case. Of greater interest are how the statics are changed by introducing a small amount of foresight: As an application of the theory, we study conditions under which increasing foresight (improving rationality as it were) can lead to either increases or decreases in welfare.

Paradoxically, we conclude that in the "usual" case increasing foresight (or rationality) will typically lower the utility of all players. By trying to control the learning of other players about their own play, players will typically be led to undertake actions even more inefficient than the static Nash equilibrium.

2. The Model

We wish to examine the behavior of players who have limited foresight, but are not completely myopic. We focus on the simplest setting in which it is possible to study this issue. Our setting is that of a simultaneous move *n* person game in which each player *i* controls an action $x_i \in [0,1]$. The profile of actions by all players is denoted by *x*. The assumption that actions are continuous is important because we will need opposing players actions to adjust continuously in order to introduce extrapolation; the assumption that they are one-dimensional avoids complications in distinguishing between different learning models such as those that give rise to the replicator, and fictitious play. One interpretation, but not as we shall see the most interesting, is that the action is a probability of playing one of two pure strategies. Whether x_i is actually observed is left deliberately vague, as the level of abstraction in the modeling is such that it makes no formal difference. However, conceptually, it is useful to think of x_i as observable only indirectly through a noisy proxy.

The flow of utility received by player *i* at a moment of time is denoted by $u_i(x)$, and reflects the entire profile of player actions. We assume that u_i is a smooth function. In the mixed strategy case (with two pure actions) u_i is linear in each argument. It is useful to recall at this point the continuous time *best response dynamic*. This says that each player's action is continuously adjusted in the direction that increases utility:[2]

$$\dot{x}_i = D_i u_i(x).$$

[2] It is possible also to include a multiplicative rate of adjustment parameter, but it is equally easy to imagine that the rate of adjustment is incorporated directly into utility.

Because each player controls only a single action, this dynamic is the same as the replicator dynamic from evolutionary game theory. It satisfies both the myopic property examined by Swinkels [1993] and the convex monotonicity of Hofbauer and Weibull [1995]. Basically, with a single action per player, any sensible model of learning in continuous time must have the property that each individual player moves in the direction of increasing utility. If we allow each player to control several actions, the situation is more ambiguous, and this issue is largely irrelevant to the discussion here, and is discussed at some length, for example, in Fudenberg and Levine [1996].

The crucial feature here is that actions adjust only gradually to the optimum (in contrast to the discrete time best response dynamic, where they adjust all at once). This is a commonly assumed feature in models of learning and evolution, such as for example, Kandori, Mailath and Rob [1993], Young [1993], and so forth. There are two basic justifications for this model, depending on whether we interpret x_i and an individual strategy (mixed perhaps) or a population aggregate. The most common model is the "evolutionary" model of a population aggregate, but we are interested in the ability of one player to learn about other players learning about him, and this is not relevant in a large population. When x_i is interpreted as an individual strategy there are two basic learning models that give rise to this type of gradual adjustment: the fictitious play model and the stimulus response model of Borgers and Sarin [1994] or Er'ev and Roth [1994]. For more discussion of the connection between these models see Cheung and Friedman [1994] or Fudenberg and Levine [1996]. In fictitious play response is gradual because the state variable is a time average that can adjust slowly. However, this model is appropriate in the current setting only in the case in which x_i is a mixed strategy; with a continuous action space, we cannot identify a state variable consisting of opponents average past

frequency of play with the opponents action. In the stimulus response model, it is simply assumed on the basis of observed behavior that adjustment to stimuli is gradual. This is the model most suitable to the current setting.

The simple best response dynamic implicitly assumes that player care only about the flow of utility. In an economic setting, they ought to care not about the flow of utility but the stock, or average present value of utility $\rho\int_0^\infty u_i(t)e^{-\rho t}dt$. Let $v_i(x)$ denote player i's perception of the average present value of utility at the current profile of actions.[3] Then the corresponding best response dynamic is

$$\dot{x}_i = D_i v_i(x).$$

If players are ignorant of opponents' future play, how can they compute average present value? If they are boundedly rational, one way to do this is to extrapolate opponent's current play. The most naïve form of extrapolation is to assume that opponent's future play will remain fixed at the current level. Then

$$v_i(x) = \rho\int_0^\infty u_i(x)e^{-\rho t}dt = u_i(x).$$

This leads to the standard best response dynamic introduced above.

Alternatively, if players are a bit more sophisticated, they may extrapolate not only current levels of utility, but also rates of change. This leads to

$$\begin{aligned} v_i(x) &= \rho\int_0^\infty \left[u_i(x) + \sum_{j\neq i} D_j u_i(x)\dot{x}_j t\right]e^{-\rho t}dt \\ &= u_i(x) + \rho^{-1}\sum_{j\neq i} D_j u_i(x)\dot{x}_j \end{aligned}$$

[3] Implicitly we are assuming that players are using *strong Markov strategies*, that is, that each player choose actions depending only on the current profile so that it indeed sensible to make inferences about future play of opponents based on these levels.

Note that we omit player i's own action from the rate of change calculation, since this is under his own control.[4] If the discount factor ρ^{-1} is low enough, then this is also the same as the standard best response dynamic above, which is one reason that play in the standard best response dynamic is called "myopic." Notice also that this model is the "stock-flow" model: a player has utility that depends both on a stock x and on a flow $\dot{x}$. This formulation is more primitive than that studied by Clower [1954], because utility here is necessarily linear in the flow. Our plan of research in this paper is to study this stock-flow model in the case in which the discount factor ρ^{-1} is not zero, but is relatively small, to discover what changes in the conclusions of the standard best response/replicator dynamic result.

[4] This is not exactly an application of the maximum principle, since we are not dealing with an optimization problem, but it is easy to check that if we include player i's own rate of change in this equation, we may solve for approximately the same equation written here.

3. An Approximation

The basic equation of motion is

$$\dot{x}_i = D_i v_i(x) = D_i u_i(x) + \rho^{-1} \sum_{j \neq i} \left[D_{ij} u_i(x) \dot{x}_j + D_j u_i(x) D_i \dot{x}_j \right].$$

The key fact is that the right hand side of the equation for $\dot{x}$ depends on both $\dot{x}$ and its derivative with respect to the levels of actions. That is, this system is in effect an equilibrium system, since player *i*'s adjustment depends on player *j*'s adjustment. When we calculate out the derivative of $\dot{x}$

$$D_k \dot{x}_i = D_{ki} u_i(x) + \rho^{-1} D_k \sum_{j \neq i} \left[D_{ij} u_i(x) \dot{x}_j + D_j u_i(x) D_i \dot{x}_j \right]$$

we see that mathematically this system is a system of partial differential equations determining in equilibrium the unknown function $\dot{x}(x)$.

In general this system of partial differential equations is quite difficult to study, as well as posing the conceptual problem about whether it makes sense to solve an equilibrium system in order to study disequilibrium behavior. However, when the discount factor ρ^{-1} is small, so that players are myopic, but not completely so, it is possible to find an approximate solution to this system that also solves many of the conceptual difficulties. Basically this approximation comes from truncating the recursion of learning about learning after a finite number of stages. Specifically, if we substitute the expressions for $\dot{x}, D\dot{x}$ back into the expression for $\dot{x}$, we drop the terms of order ρ^{-2} which correspond to learning about learning about learning, but keep the terms of order ρ^{-1} which correspond to learning about learning. If in fact ρ^{-1} is sufficiently small, this will be a good approximate solution to the partial differential equation system.

Specifically, we compute

$$\dot{x}_i = D_i v_i(x) = D_i u_i(x) + \rho^{-1} \sum_{j \neq i} \left[D_{ij} u_i(x) \dot{x}_j + D_j u_i(x) D_i \dot{x}_j \right] \approx D_i u_i(x)$$

$$D_k \dot{x}_i = D_{ki} u_i(x) + \rho^{-1} D_k \sum_{j \neq i} \left[D_{ij} u_i(x) \dot{x}_j + D_j u_i(x) D_i \dot{x}_j \right] \approx D_{ki} u_i(x) .$$

If we then substitute into the expression for $\dot{x}_i$ these approximations, we find the approximate solution

$$\dot{x}_i \approx D_i u_i(x) + \rho^{-1} \sum_{j \neq i} \left[D_{ij} u_i(x) D_j u_j(x) + D_j u_i(x) D_{ij} u_j(x) \right].$$

This system of dynamical equations is not the solution to an equilibrium system; rather each player correctly anticipates the impact he will have on his opponents' myopic learning (that is, learning about levels), but not the impact he will have on his opponents sophisticated learning. This is consistent with the idea of bounded rationality. This system is an approximate solution to the original system of partial differential equations in the sense that if the higher derivatives of u_i are all bounded, then the error when we substitute this "solution" into the original system is of order ρ^{-2}.

Note that in the one player case we have

$$\dot{x}_i = D_i u_i(x)$$

which is the ordinary continuous time best-response dynamic. The control variable is adjusted upwards in the direction of increasing utility; steady states correspond either to maxima or minima of the utility function, and only maxima are stable. Note also that this is essentially the same dynamic studied in the traditional stock/flow investment model introduced in Clower [1954].

4. Review of the Myopic Case

In the limiting case of a zero discount factor $\rho^{-1} = 0$, we have the perfectly ordinary continuous time best response dynamic

$$\dot{x}_i = D_i u_i(x).$$

Steady state of this system satisfy $\dot{x}_i = D_i u_i(x) = 0$ and correspond exactly to Nash equilibria of the static game provided that utility is concave in each player's own action, that is, $D_{ii} u_i(x) < 0$.

This system may converge to a steady state or may cycle, or even exhibit chaotic behavior. If we focus on interior equilibria (in case x_i is a mixture over two pure strategies, this corresponds to a completely mixed equilibrium), then the condition for dynamic stability is that the eigenvalues of the matrix

$$\left[D_{ij} u_i(x)\right]_{i,j=1,\ldots n}$$

should have negative real parts. In the special case of x_i a mixture over two pure strategies $D_i u_i$ does not depend on x_i since u_i is linear in x_i, so the diagonal of $\left[D_{ij} u_i(x)\right]_{i,j=1,\ldots n}$ is zero, and so the matrix has zero trace. It follows that in this case that the steady state cannot be both stable and hyperbolic.

5. Steady State Analysis in the Symmetric Case

To more clearly understand the dynamics resulting from learning about learning, observe first that as ρ^{-1} goes to zero, so we approach the myopic case, steady states of the system will approach those of the myopic system. Moreover, if the steady states of the myopic system are hyperbolic, so robust to perturbations, then the stability properties of steady states will approach those of the myopic system. The interesting question, therefore, lies in knowing in what direction the steady state will move due to learning about learning. To understand this more clearly, we focus on the symmetric case.

In the case in which all players are identical, including in initial condition, the dynamic system can be simplified to a single dimensional dynamical system

$$\dot{x}_i \approx D_i u_i(\vec{x}_i) + \rho^{-1}(n-1)\left[D_{ij}u_i(\vec{x}_i)D_i u_i(\vec{x}_i) + D_j u_i(\vec{x}_i)D_{ij}u_i(\vec{x}_i)\right]$$

where $\vec{x}_i = (x_i, x_i, \ldots, x_i)$, and j is any index not equal to i. At a steady state $\hat{x}_i(\rho^{-1})$, $\dot{x}_i = 0$. We may differentiate this condition to find

$$D_\rho \hat{x}_i(0) = -\left[D_{ii}u_i(\vec{x}_i) + (n-1)D_{ij}u_i(\vec{x}_i)\right]^{-1}(n-1)D_j u_i(\vec{x}_i)D_{ij}u_i(\vec{x}_i).$$

To understand better what this means, consider how utility of a player changes if all players increase their action

$$Du_i(\vec{x}_i) = D_i u_i(\vec{x}_i) + (n-1)D_j u_i(\vec{x}_i).$$

Making use of the condition at a myopic equilibrium that $D_i u_i(\vec{x}_i) = 0$, this may be written as

$$Du_i(\vec{x}_i) = (n-1)D_j u_i(\vec{x}_i).$$

Combining this with the previous result, we can find the change in utility with respect to a small change in the discount factor ρ^{-1} away from zero $Du_i(\vec{x}_i)D_\rho\hat{x}_i(0)$. Consequently, steady state utility will increase if

$$-\left[D_{ii}u_i(\vec{x}_i)+(n-1)D_{ij}u_i(\vec{x}_i)\right]^{-1}D_{ij}u_i(\vec{x}_i)>0,$$

and decrease otherwise.

Next, we should focus attention only on steady states that are stable, and in particular are stable in the myopic case. Since the myopic dynamic is

$$\dot{x}_i = D_i u_i(\vec{x}_i)$$

the corresponding stability condition is $D\dot{x}_i < 0$, or

$$D\dot{x}_i = D_{ii}u_i(\vec{x}_i)+(n-1)D_{ij}u_i(\vec{x}_i)<0.$$

It follows that small changes in the discount factor ρ^{-1} away from zero will increase steady state utility near a stable steady state if

$$D_{ij}u_i(\vec{x}_i)>0,$$

and decrease it otherwise. Moreover, if we bound the effect that players can have on their own and other players utility (that is, $D_{ii}u_i, D_{ij}u_i$), and consider a large number of players, it is clear that the stability condition implies $D_{ij}u_i(\vec{x}_i)$ must either be non-positive, or at least approach zero. From this we reach the paradoxical conclusion that unless there are strategic complementaries, the effect of increasing the patience of the players will be to decrease the utility of all players.

6. Learning about Learning

At this stage we have to ask whether any computationally feasible learning process could lead to the inferences driving the stock flow model of learning. We now examine that question in more detail in the context of a simple example. We examine the simplest interesting case of a symmetric game with two players, and $u_i(x_i, x_j) = cx_i x_j$. Throughout the discussion we assume that both players observe their opponents choices after the fact.

Our basic adjustment equation can be written

$$\begin{aligned}\dot{x}_i &= D_i u_i(x) + \rho^{-1}\left[D_{ij}u_i(x)\dot{x}_j + D_j u_i(x) D_i \dot{x}_j\right] \\ &= cx_j + \rho^{-1}\left[c\dot{x}_j + cx_i D_i \dot{x}_j\right]\end{aligned}$$

It is convenient now to define $y_i = \log x_i$, and convert this equation into logarithmic form

$$\begin{aligned}\dot{x}_i / x_i &= c(x_j / x_i) + \rho^{-1}\left[c(x_j / x_i)\dot{x}_j / x_j + cx_i(x_j / x_i) D_i \dot{x}_j / x_j\right] \\ \dot{y}_i &= c\exp(y_j - y_i) + \rho^{-1}\left[c\exp(y_j - y_i)\dot{y}_j + c\exp(y_j - y_i) D_i \dot{y}_j\right]\end{aligned}$$

Given symmetry, we may assume that y_i, y_j are close, and introduce the approximation

$$\dot{y}_i \approx c(1 + y_j - y_i)\left[1 + \rho^{-1}\left[\dot{y}_j + D_i \dot{y}_j\right]\right].$$

Moving to discrete time, and assuming as is usual in the learning literature that players learn through random experimentation or mutation, we may write the basic dynamical equation

$$y_i(t) = y_i(t-1) + c(1 + y_j(t-1) - y_i(t-1))\left[1 + \rho^{-1}\left[y_j(t-1) - y_j(t-2) + D_i \dot{y}_j(t-1)\right]\right] + \varepsilon_i(t)$$

.

Here the critical element is $D_i\dot{y}_j(t-1)$ which must somehow be estimated from the available data on the history of play. This amounts to estimating the dependence of $y_j(t-1)-y_j(t-2)$ on $y_i(t-2)$ (presuming that players are bright enough to realize that $y_i(t-1)$ cannot possibly have any impact on $\dot{y}_j(t-1)$. Since $y_j(t-1)-y_j(t-2)$ depends also on time, we can think of carrying out this inference by, for example, regressing $y_j(t-1)-y_j(t-2)$ on a constant and $y_i(t-2)$.

The actual relationship from the dynamical equations is

$$y_j(t-1)-y_j(t-2)=c(1+y_i(t-2)-y_j(t-2))\left[1+\rho^{-1}\left[y_i(t-2)-y_i(t-3)+D_j\dot{y}_i(t-2)\right]\right]+\varepsilon_j(t-1)$$

The key point is that if the process is non-explosive to a good approximation when ρ^{-1} is small (as in the continuous time case studied above), then

$$y_j(t-1)-y_j(t-2)=c-cy_j(t-2)+cy_i(t-2)+\varepsilon_j(t-1).$$

Given this that the data generated by the process will approximately satisfy this relationship, any reasonable procedure for estimating the dependence of $\dot{y}_j(t-1)$ on $y_i(t-2)$, $y_j(t-2)$ and a constant will approximately yield the approximately correct answer

$$D_i\dot{y}_j(t-1)\approx c.$$

This, of course, leads back (approximately) to the type of stock flow dynamics we studied in the previous sections.

Key to this result is the non-explosivity of the process. To get a handle on this, we first observe that the levels grow at a constant rate, and that the Jacobian matrix of the first difference equation for $y_j(t-1)$ is

$$\begin{bmatrix}1-c & c\\ c & 1-c\end{bmatrix}+\rho^{-1}(\text{other terms})$$

For ρ^{-1} sufficiently small, the eigenvalues will be approximately equal to 1 and 1-2*c*. The eigenvalue 1 corresponds to the eigenvector (1,1) pointing in the direction of constant linear increase, the eigenvalue 1-2*c* corresponds to the eigenvector (1,-1) measuring departures from the 45 degree line. This implies for $c > 0$ that departures from symmetry are self-correcting over time.

Inspecting the dynamical equation more closely when $y_i = y_j$, we find

$$\{y_i(t) - y_i(t-1)\} = c + c\rho^{-1}\left[\{y_i(t-1) - y_i(t-2)\} + D_i\dot{y}_j(t-1)\right] + \varepsilon_i(t)$$

Clearly, if ρ^{-1} is small enough this will be stable.

References

Borgers, T. and R. Sarin [1994]: "Learning through Reinforcement and the Replicator Dynamics," Tilburg.

Cheung, Y. and D. Friedman [1994]: "Learning in Evolutionary Games: Some Laboratory Results," Santa Cruz.

Clower, Robert W. [1954]: "An Investigation Into the Dynamics of Investment," *American Economic Review*, 44, 64-81.

Er'ev, I. and A. Roth [1994]: "On The Need for Low Rationality Cognitive Game Theory: Reinforcement Learning in Experimental Games with Unique Mixed Strategy Equilibria," University of Pittsburgh.

Fudenberg and Levine [1996], *Theory of Learning in Games*, forthcoming, MIT University Press.

Hofbauer, J. and J. Weibull [1995]: "Evolutionary Selection against dominated strategies," mimeo, Industriens Utredningsintitut, Stockholm.

Kandori, M., G. Mailath and R. Rob [1993]: "Learning, Mutation and Long Run Equilibria in Games," *Econometrica,* 61, 27-56.

Swinkels, J. [1993]: "Adjustment Dynamics and Rational Play in Games," *Games and Economic Behavior* 5, 455-484.

Young, P. [1993]: "The Evolution of Conventions," *Econometrica*, 61, 57-84.