LEARNING BY DOING AND THE VALUE OF OPTIMAL EXPERIMENTATION

Volker Wieland[1]
Monetary Affairs Division, Mail Stop 31
Board of Governors of the Federal Reserve System
Washington, DC 20551

Abstract

Research on the implications of learning-by-doing has typically been restricted to specifications of the agent's decision problem for which estimation and control can be treated separately. Recent work has provided the limit properties of beliefs and actions for learning problems under more general conditions, for which experimentation is an important aspect of optimal control. However under these conditions the optimal policy cannot be derived analytically, because Bayesian learning introduces a nonlinearity in the dynamic programming problem. This paper utilizes numerical algorithms to characterize the optimal policy function for such a general learning-by-doing problem. In contrast to previous work on calculating such policies, we find that the optimal policy incorporates a substantial degree of experimentation under a wide range of initial beliefs about the unknown parameters. Dynamic simulations indicate that optimal experimentation dramatically improves the speed of learning and the stream of future payoffs. Furthermore dynamic simulations reveal that a policy, which separates control and estimation and does not incorporate experimentation, frequently induces a long-lasting bias in the control and target variables. While these sequences tend to converge steadily under the optimal policy, they frequently exhibit non-stationary behavior when estimation and control are treated separately.

JEL No.: C44,C60, D81, D82
Key Words: optimal Bayesian learning, learnin-by-doing and experimentation, optimal control with unknown parameters, dynamic programming

[1] I would like to thank John Taylor, Ken Judd, Michael Horvath, Ronald McKinnon, Andrew Levin, Athanasios Orphanides, David Wilcox and Tom Sargent for many helpful comments and suggestions. In addition seminar participants at Stanford, the Federal Reserve Board, the University of Illinois, the European Economic Association and the World Econometric Society Meeting deserve thanks for stimulating discussions. All remaining errors are my own. The views expressed in this paper are solely the responsibility of the author and should not be interpreted as reflecting those of the Federal Reserve Board.

1. Introduction

Economic agents very rarely have complete knowledge of all parameters that affect their payoffs. Yet usually they are able to learn more about these parameters by observing the outcome of their own actions. That is, the agents' current actions not only determine current payoffs but also provide useful information which could improve future actions and yield superior future payoffs. The standard approach to this problem is to treat estimation and control separately. Its recipe for the agent is to first estimate the unknown parameters and then treat these values as certain when choosing an action to maximize payoffs. This paper shows that the standard approach, which neglects the information-producing effect of current actions, can be far from optimal. The paper deals with specifications of the economic environment where agents are faced with a trade-off between actions yielding a high current payoff and actions which yield a lower current payoff but superior information content.

So far, most research on economic implications of learning by doing[1] and on-the-job training has been restricted to the few specifications of the learning problem, for which it is optimal to treat control and estimation separately. For more general cases, the appropriate extent of experimentation is quite difficult to determine. However, substantial progress has been made concerning the limit properties of beliefs and actions in general learning by doing problems through recent work by Aghion, Bolton, Harris and Jullien (1991), Easley and Kiefer (1988) and Kiefer and Nyarko (1989).[2] For example, Kiefer and Nyarko (1989) prove that posterior beliefs and associated actions in linear controlled regressions with Bayesian learning converge in the limit, but they also show that limit beliefs need not coincide with the true parameter values.

This paper characterizes the optimal policy for general learning by doing problems that take the form of linear controlled regressions. It assesses the importance of experimentation and its implications for the time series behavior of economic observables. Even for relatively simple

[1]For example recent work on technological innovations by Jovanovich and Nyarko (1994) and (1995) and on agricultural production by Foster and Rosenzweig (1995), which constitute major advances with respect to other aspects of learning-by-doing still avoids the issue of optimal experimentation.

[2] Related economic applications have dealt with the monopolist's problem under unknown demand, McLennan (1984), Kiefer (1989) and Trefler (1993) and with monetary policy and transition, Bertocci and Spagat (1993) and Balvers and Cosimano (1994).

specifications of parameter uncertainty this is not an easy task because learning introduces a nonlinearity into the dynamic programming problem and typically precludes analytical solutions. Characterizing the value and extent of experimentation is necessary, however, for answering many of the questions that arise in economic applications of learning by doing.

This paper provides answers to three such questions:

(i) Is experimentation important, or in other words, does the passive learning policy that separates control and estimation differ significantly from the optimal learning policy for a non-negligible range of beliefs?

(ii) To what extent does experimentation accelerate the speed of learning and guarantee that the agent learns the true parameter values?

(iii) Could the behavior of economic agents that are learning about their environment be the source of non-stationarities in economic data?

The learning by doing problem considered hereafter is equivalent to the problem of controlling a linear regression process that has two unknown parameters. The agent controls the independent variable in order to influence the dependent variable which enters his payoff function. At the same time, the agent continuously updates his beliefs about the unknown parameters according to Bayes rule. The myopic policy, which treats estimation and control of the stochastic process separately, can be easily derived analytically, while the optimal learning policy has to be approximated numerically.

The first step in solving this problem is to explicitly model the agent's beliefs about the unknown parameters. As an illustrative example, I focus on a variant of the two-armed bandit problem studied by McLennan (1984) and Kiefer (1989), in which the agent is faced with two possible sets of parameter values. This specification is not realistic, but significantly simplifies the dynamic programming problem, since the state can be summarized by a single variable: the probability that one set of values is the correct one. Then I turn to the more useful problem of controlling a linear regression, where from the agent's perspective the values of the unknown parameters can be any real number. This specification models the agent's beliefs as a bivariate normal distribution, which implies a complex dynamic programming problem with five state variables: the two means, the two variances and the covariance.

Prescott (1972) was the first[3] to use numerical techniques to approximate the optimal policy as a function of beliefs for the case of a simple regression with unknown slope. He found little difference between myopic and optimal policies except with very high parameter uncertainty i.e. when the slope estimate is insignificant. These results are reversed here. This paper detects sizeable differences between optimal and myopic policies, or in other words a sizeable extent of optimal experimentation, in a problem with two unknown parameters. Even though there is only one more unknown parameter the problem is considerably more complicated than Prescott's because the number of state variables increases from one to five and because there exist multiple limit beliefs and policies.

I obtain several new results which are relevant to the theory of optimal learning, and furthermore have important implications for time series studies of economic observables generated by agents who are learning about their economic environment. I find that optimal experimentation plays an important role in speeding up learning and can significantly improve expected future payoffs, because it has a high reward for a large range of beliefs about the unknown parameters. This result is due to the existence of multiple uninformative actions that would reinforce an incorrect belief even in the limit, but would still be chosen by an agent who does not experiment. I also find that these incorrect belief and action pairs are associated with non-differentiabilities in the value function and discontinuities in the policy function. Finally, the dynamic simulations show that optimal learning generates time series that greatly differ form those generated under passive learning. While passive learning frequently results in persistent biases in beliefs, actions and observations of agents learning about their economic environment, this happens very rarely if these agents experiment optimally.

The next section presents the learning by doing problem, reviews relevant convergence results available in the literature and specifies the illustrative example as well as the general case. Section 3 reports numerically approximated value and policy functions which characterize the value and optimal extent of experimentation. Section 4 studies the time series behavior of economic observables under passive versus optimal learning and Section 5 concludes.

---

[3]Taylor (1974) and Taylor and Anderson (1976) conducted simulation studies to investigate convergence of parameter estimates under myopic policies which can be derived analytically. Kendrick (1981) and others used numerical techniques, called dual control, to approximate active learning policies along given sample paths.

2. Learning By Doing when Separating Control and Estimation is not Optimal

2.1. The Decision Problem

This paper uses an information-theoretic model of learning-by-doing, which is equivalent to the problem of controlling a linear regression process with unknown parameters. The model has a single decision maker who observes a signal y at time t, which depends on his own action x, the unknown parameters of the regression process and a random independent shock.

(1)

$$y_t = \alpha + \beta x_t + \epsilon_t$$

The agent's payoffs are a function of the action and the signal.[4] To the extent that the agent cares about future payoffs, he must also take into account the information-accumulating effect of his current actions. As a model of learning by doing this is considerably more general than earlier treatments in the literature, because it is not optimal to address the estimation and control aspects separately.

The sequence of events is as follows: based on his beliefs about the unknown parameters at time t, the agent chooses a value for the control x_t. Then a shock ϵ_t occurs and a new observation on y_t becomes available. Before choosing next period's control x_{t+1}, the agent updates his estimates of α and ß using the new information contained in (x_t, y_t). ϵ_t is assumed to be normally distributed with mean zero and known variance σ^2.[5] The agent's beliefs with respect to the parameters $(\alpha,ß)$ are modelled as a probability distribution $p(\alpha,ß|\theta_t)$ where θ_t is the vector of parameters characterizing the distribution p(.). The prior $p(\alpha,ß|\theta_{t=0})$ constitutes the agent's initial belief about α and ß. When a new observation on y_t becomes available, these beliefs are

[4] This is the framework studied by Kiefer and Nyarko (1989). which differs somewhat from Aghion, Bolton, Harris and Jullien (1991). Since the signal is a function of the action, the parameters and the shock, but the agent's payoff is a function of the action and signal only, he must consider the direct effect of his action on his signal, its direct effect on his payoff and its indirect effect (via the signal) on the payoff. In the framework of Aghion et al., the payoff obtained in a given period and the signal observed, depend on the action taken, the underlying parameter of interest and a shock. So the optimizing agent faces a simple trade-off between choosing his action to maximize his payoff and choosing it to obtain the best possible signal. They find that local properties of the payoff function (such as smoothness) are crucial in determining whether the agent eventually attains the true maximum payoff or not. There is some similarity to the results obtained in this paper, which finds that kinks in the value function are related to convergence to the true maximum payoff.

[5] For all the results obtained in this paper the error variance has been set to one. However it is straightforward to investigate the sensitivity of the results to the size of σ^2 or to a misperception of the agent about the size of σ^2.

updated according to Bayes rule

(2)

$$\theta_{t+1} = B(y_t, x_t, \theta_t)$$

where B(.) is the Bayes operator. Equation (2) shows that the current action x_t not only affects the current realization of y but also next period's estimates of the unknown parameters and through that next period's realizations of x and y. Thus a forward-looking agent may not find it optimal to solve the control and estimation problems separately, and rather take advantage of the information-accumulating effect of current actions.

The agent's objective is characterized by a one-period payoff function U(y,x). Given this payoff function one obtains expected one-period reward $R(x,\theta)$ by taking expectations with respect to α,ß and ϵ[6]:

(3)

$$R(x,\theta) = \int_R \int_{R^2} U(\alpha + \beta x + \epsilon, x)\, p(\alpha, \beta \,|\, \theta)\, q(\epsilon)\, d\alpha\, d\beta\, d\epsilon$$

An agent which treats the control and estimation problems separately will choose a control policy which maximizes expected one-period reward based on current parameter estimates. Once y_t is observed, these parameter estimates are updated and next period's control x_{t+1} is adjusted appropriately. The behavior resulting from such a choice is myopic and implies passive learning about the unknown parameters. A policy is stationary if it can be represented by a function $H(\theta_t)$ which selects an action x_t based solely on the current state θ_t. In other words, it constitutes a closed-loop control. Once the distribution representing the agent's beliefs has been specified the stationary myopic policy $H^{my}(\theta_t)$ can be easily derived analytically. However it is much harder to obtain the stationary optimal learning policy $H^{opt}(\theta_t)$ which maximizes the discounted sum of expected current and future rewards:

(4)

$$\underset{[x_t]_{t=0}^{\infty}}{\text{Max}}\; E^{[H,\theta_0]} \left[\sum_{t=0}^{\infty} \delta^t R(x_t, \theta_t) \;\middle|\; \theta_0 \right]$$

Note that the expectations operator is not only induced by the prior belief θ_0 but also by the

---

[6] The distribution $q(\epsilon)$ is assumed to be known to the agent.

policy function $H(\theta)$. If α and ß were known, this would be a static problem and (4) could be maximized by a sequence of one-period optimal actions. But with $(\alpha,ß)$ unknown, beliefs change over time and form an explicit link between present and future periods. Estimation and control cannot be separated because future beliefs $(\theta_{t+j}, j=1,2,...)$ depend on the whole history of actions $(x_{t+j-1}, j=1,2,...)$ and thus on the function $H(\theta)$. The ensuing difficulty in calculating the optimal learning policy becomes clear once the problem is brought into a dynamic programming framework. For this purpose the beliefs θ_t are interpreted as state variables. The value function $V(\theta)$ defines the supremum of (4). It is achieved by the optimal learning policy $H^{opt}(\theta)$ and must satisfy the functional equation:

(5)

$$V(\theta) = \underset{x}{\text{Max}} \left[R(x,\theta) + \delta \int V(B(x,y,\theta)) f(y|\, x,\theta)\, dy \right]$$

where next period's beliefs θ' have been replaced by the Bayes operator B(.). $f(y|x,\theta)$ denotes the predictive distribution of y which depends on x, the distribution of the error term q(.) and the beliefs p(.). The two terms on the right-hand side of (5) characterize the potential trade-off between current control and estimation. The first term is the current expected reward while the second term represents the expected improvement in future payoffs due to improved information about the unknown parameters. It arises because the choice of x and the resulting observation on y add to the agent's stock of information.

Kiefer and Nyarko (1989) prove that a stationary optimal policy exists and that the value function $V(\theta)$ is continuous and satisfies the functional equation (5).[7] Policy and value functions can be obtained by iterating over the functional equation starting with an initial guess about the form of V(.). However, even if a quadratic loss function is used as an initial guess, the integration in (5) usually cannot be carried out analytically to obtain a closed form solution. This difficulty arises because B(.) is a nonlinear function of y and x, and x also appears in the predictive distribution of y. There are many examples, including the cases addressed in this

[7] They use standard dynamic programming methods and show that Blackwell's sufficiency condition - monotonicity and discounting - are satisfied. Therefore (5) has a fixed point in the space of continuous functions, which is the value function $V(\theta)$.

paper, which do not admit analytic solutions for optimal learning policies, even though the unknown process is linear and preferences are quadratic. Whether numerical approximation is feasible and how difficult it is, depends on the specification of the agent's beliefs and learning process. In section 3 I present results for two specifications, an illustrative example and a general case. All results are based on a quadratic payoff function

(6)
$$U(y,x) = -(y-y^*)^2 - \omega x^2$$

where y^* is the desired target level and the weight $\omega \geq 0$. This coincides with the input-target model used in studies of learning by doing and technological change. However the numerical algorithm admits other functional forms and the ultimate choice should depend on the specific economic application of learning by doing.

2.2. Convergence of Beliefs and Actions

The essential contribution of Easley and Kiefer (1988) and Kiefer and Nyarko (1989) (KN hereafter) is to prove convergence of beliefs and policies under minimal distributional assumptions and to characterize the set of possible limit beliefs and policies for the case of a simple linear regression. Standard convergence results are not applicable, because along any sample path for which parameter estimates converge the sequence of actions is also converging. Thus, if actions converge rapidly they may not generate enough information for identifying the unknown coefficients and the limit distribution representing the agent's limit beliefs with respect to the unknown parameters needs not to be centered on the true parameter values.

KN show that the posterior process of beliefs θ_t converges to a limit $\overline{\theta}$ for any multiple linear regression process.[8] However, this convergence result says nothing in particular about the limit belief and whether it is correct. For the case of a simple regression with a single control variable and known error distribution, KN show that: (i) if the policy x_t does not converge then the posterior process converges and the limit is centered at the true values, which means complete

[8] The distributional assumptions made in this paper are only necessary to solve for an optimal policy. KN prove convergence of the posteriors by an application of the martingale convergence theorem without restricting beliefs to any conjugate families. The posterior process can be shown to be a martingale, in other words, the agent does not expect his beliefs to change in any predictable manner. For the two cases studied in this paper based on normal beliefs it can easily be ascertained that the Bayesian updating equations in (11) and (19) are martingales.

learning occurs; (ii) if x is fixed, the agent learns the mean of y corresponding to that value of x, which means even if the limit belief and action pair are incorrect the agent always learns the true mean of y; and (iii) if the reward function is strictly concave, the limit policy $\overline{x}$ exists and corresponds to the policy maximizing single period reward given limit beliefs $\overline{\theta}$.

KN point out that incomplete learning can be optimal, or in other words, that there are multiple limit beliefs, some of which are not point mass at the true coefficient values but are associated with optimal limit actions. The underlying intuition is that since learning is costly, an agent may choose actions which are suboptimal given the truth but optimal given his subjective beliefs. The above results however provide three properties that are common to all possible limit belief and action pairs $(\overline{\theta},\overline{x})$, whether incorrect or not:

belief invariance: limit beliefs must be invariant to the Bayesian updating rule which means that the updating rule must have a fixed point at any limit belief and action pair.

(7)
$$\overline{\theta} = B(\overline{x},\overline{\theta},\epsilon)$$

prediction: in the limit the agent learns the mean of y associated with the policy $\overline{x}$.

(8)
$$E[\alpha|\,\overline{\theta}] + E[\beta|\,\overline{\theta}]\,\overline{x} = \alpha + \beta\,\overline{x}$$

optimization: the limit action $\overline{x}$ maximizes expected reward given limit beliefs $\overline{\theta}$.

(9)
$$\operatorname*{Max}_{x}\; E\left[\,U(\alpha + \beta\,x + \epsilon, x)\,|\;\overline{\theta}\,\right]$$

Once the agent's beliefs and learning process have been specified, these three properties can be used to characterize the set of possible limit belief and action pairs and to determine whether there are any incorrect limit beliefs among these.

2.3. Specifying the Agent's Beliefs and Learning Process

Throughout the remainder of the paper, the illustrative example is presented parallel to the general specification for which the two unknown parameters can take any real value. This makes it possible to discuss most of the relevant issues in the example while avoiding initially the complexity arising from multiple state variables and multiple possible limit beliefs.

Illustrative Example

There are two sets of parameter values (α_1,β_1) and (α_2,β_2), of which the first is considered the correct one. The vector θ_t which characterizes the agent's belief has only one element, namely the probability p_t that $(\alpha,\beta)=(\alpha_1,\beta_1)$, which then constitutes the single state variable of the dynamic programming problem[9]. Table 1 summarizes the fully specified dynamic programming problem for this example.[10]

Table 1. Specification of the Illustrative Example

predictive distribution of y

$$f(y|x,p) = (2\pi)^{-1/2}\left(p_t e^{-1/2(y_t-\alpha_1-\beta_1 x_t)^2} + (1-p_t)\, e^{-1/2(y_t-\alpha_2-\beta_2 x_t)^2}\right)$$

updating equation

$$p_{t+1} = B(p_t,x_t,y_t) = \frac{p_t e^{-1/2(y_t-\alpha_1-\beta_1 x_t)^2}}{p_t e^{-1/2(y_t-\alpha_1-\beta_1 x_t)^2} + (1-p_t)\, e^{-1/2(y_t-\alpha_2-\beta_2 x_t)^2}}$$

expected one-period reward

$$R(x_t,p_t,\alpha_1,\alpha_2,\beta_1,\beta_2) = -[1+p_t\alpha_1^2+(1-p_t)\alpha_2^2+2x_t(p_t\alpha_1\beta_1+(1-p_t)\alpha_2\beta_2)+(p_t\beta_1^2+(1-p_t)\beta_2^2)x_t^2]$$

myopic policy

$$x^{my} = H^{my}(p_t) = -\frac{p_t\alpha_1\beta_1+(1-p_t)\alpha_2\beta_2}{p_t\beta_1^2+(1-p_t)\beta_2^2}$$

functional equation

$$V(p) = \underset{x}{\mathrm{Max}}\left[R(x,p,\alpha_1,\alpha_2,\beta_1,\beta_2) + \delta\int V(B(p,x,y))f(y|\,x,p)\,dy\right]$$

---

[9] Using a specification with a single state variable has two advantages: it significantly reduces the computational costs of numerically approximating value of policy functions, and since these functions are one-dimensional they can easily be graphed and directly compared to the myopic policy and single-period reward functions. This example is a simple variation of the problem studied by Kiefer (1989) and used by El-Gamal and Sundaram (1993).

[10] The payoff function which corresponds to expected current reward is the quadratic loss function of equation (6). For simplicity the target level y^* and the weight ω are set to zero initially.

Both, passive and optimal learning policies have to be compared to the optimal action under certainty which is independent of p_t. If (α_1,β_1) are the true parameter values, it is simply

(10)
$$x^* = -\frac{\alpha_1}{\beta_1}$$

and the correct belief is $p_t=1$. Learning the true parameter values implies that parameter estimates are consistent and converge in the limit to the true values ($p_t \rightarrow 1$) while the policy converges to (10).

Using the three properties of limit belief and action pairs discussed in section 2.2. (belief invariance, prediction and optimality) one can show that there exists either one or no incorrect limit belief depending on the values of (α_1,β_1) and (α_2,β_2). Beliefs are invariant, if the updating rule in table 1 has a fixed point, which is true for p=0, p=1, and any p associated with an action $\overline{x}$ s.t.

(11)
$$\alpha_1 + \beta_1 \bar{x} = \alpha_2 + \beta_2 \bar{x}$$

(11) defines the intersection of the two curves implied by the parameter values, when the error term is omitted. It defines the single value of x, which is consistent with both curves in the limit. Any pair (p,x) which satisfies (11) also has the 'prediction' property. The' optimality' property implies that the limit action maximizes expected one-period reward based on the limit belief:

(12)
$$\bar{x} = -\frac{(\bar{p}\,\alpha_1\beta_1 + (1-\bar{p})\,\alpha_2\beta_2)}{\bar{p}\beta_1^2 + (1-\bar{p})\,\beta_2^2} \qquad \text{where } 0 < \bar{p} < 1$$

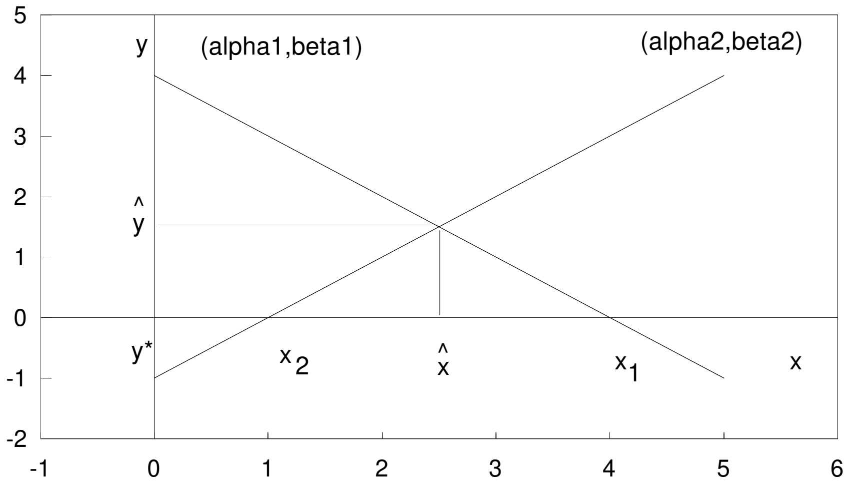
A possible incorrect limit belief and action pair consists of an incorrect but self-reinforcing belief (a belief which calls for an action x satisfying (11)), with a reinforcing action which is one-period optimal given the incorrect belief (an action x satisfying (12)).

Figure 1 shows that if the two curves have opposite slopes, there exists one such possible incorrect limit belief and action pair denoted by $(\hat{p},\hat{x})$. $\hat{x}$ corresponds to the intersection of the two curves, and therefore is an uninformative action which would reinforce any belief p.

Equation (12) determines the belief $\hat{p}$ for which $\hat{x}$ is a one-period optimal action. It defines the one-period optimal action under uncertainty as an average of x_1 and x_2, which are the optimal actions under certainty about (α_1,β_1) and (α_2,β_2) respectively. Since $\hat{x}$ lies in the interval between x_1 and x_2 as can be seen from the graph, there must exist a belief $\hat{p}$ for which $\hat{x}$ is one-period optimal. The numerical example of an incorrect limit belief and action pair depicted in the graph is $(\hat{p}=0.5,\hat{x}=2.5)$ based on values of $(\alpha_1=4,\beta_1=-1)$and $(\alpha_2=-1,\beta_2=1)$, which imply $x_1=4$ and $x_2=1$.[11]

[11] When the slopes (β_1,β_2) of the two possible curves have the same sign, there does not exist any optimal incorrect limit belief.

Figure 1 Illustrative Example - Two Possible Relationships



General Case

The previous example restricts the agent's beliefs to two possible sets of coefficient values. More realistic and more useful is a setup where the specification of beliefs allows these coefficients to take any real value. This can be achieved by modeling the agent's beliefs as a bivariate normal distribution with respect to α and ß:

(13)

$$p(\alpha,\beta\,|\,\theta_t) = N\left(a_t,b_t,\Sigma_t\right) \qquad \text{where} \quad \Sigma_t = \begin{pmatrix} v_a & v_{ab} \\ v_{ab} & v_b \end{pmatrix}_t$$

$p(\alpha,\beta|\theta_t)$ is fully characterized by the vector θ_t which contains five elements: the means (a_t,b_t) and the variances $(v_{a,t},v_{b,t})$ and covariance $v_{ab,t}$ summarized in the variance-covariance matrix Σ_t. **Table 2** summarizes the dynamic programming problem based on bivariate normal beliefs.

Table 2. Specification of the DP Problem

updating equations

$$\Sigma_{t+1} = \left[\,\Sigma_t^{-1} + X_t X_t\right]^{-1} \qquad \begin{pmatrix} a \\ b \end{pmatrix}_{t+1} = \Sigma_{t+1}\left[X_t^{\prime} y_t + \Sigma_t^{-1}\begin{pmatrix} a \\ b \end{pmatrix}_t\right]$$

expected one-period reward

$$R(x,a,b,\Sigma) = -[\,1 + a^2 + x^2(v_b + b^2 + \omega) + 2x(v_{ab} + ab)\,]$$

myopic policy

$$x^{my} = H^{my}(a_t,b_t,\Sigma_t) = -\frac{v_{ab} + ab}{v_b + b^2 + \omega}$$

functional equation

$$V(a,b,\Sigma) = \underset{x}{\text{Max}}\left[\,R(x,a,b,\Sigma) + \delta \int V(B(a,b,\Sigma,x,\alpha+\beta x+\epsilon))\,q(\epsilon)\,p(\alpha,\beta\,|\,a,b,\Sigma)\,d\alpha\, d\beta\, d\epsilon\,\right]$$

The derivation of the updating equations is shown in Zellner (1971). X_t denotes the vector $(1,x_t)$. The form of the functional equation is obtained by substituting out the dependent variable y.

Given the true parameter values α and ß, the correct belief would be $a_t=\alpha$, b_t=ß and $v_{ab}=v_a=v_b=0$. It can be shown that there exist multiple limit beliefs - all but one incorrect - which can all be associated with optimal actions. Using the three properties discussed in section 2.2 (belief invariance, prediction, optimality) the set of possible limit belief and action pairs can be characterized by the following collection of 4 equations and 3 inequality conditions which has multiple solutions.

Table 3. The Set of Possible Limit Beliefs and Actions

belief invariance	$\bar{v}_a + \bar{v}_{ab}\bar{x} = 0$ $\bar{v}_{ab} + \bar{v}_b\bar{x} = 0$
prediction	$\alpha + \beta\bar{x} = \bar{a} + \bar{b}\bar{x}$
optimality	$\bar{x} = -\dfrac{(\bar{v}_{ab} + \bar{a}\bar{b})}{(\bar{v}_b + \bar{b}^2 + \omega)^{-1}}$
semi-positive-definiteness of Σ	$v_a v_b - v_{ab}^2 \geq 0$
non-negativity of variances	$v_a, v_b \geq 0$

The updating equations of the bivariate normal specification have a fixed point for any collection of (x,a,b,Σ) which satisfies the first two equations reported in table 3 and does not violate any non-negativity conditions with respect to the variances and the determinant of Σ. One-period optimality of limit actions implies that the set of possible limit belief and action pairs also depends on the parameters of the one-period payoff function U(.). For $\omega>0$ and $y^*=0$, optimality implies the fourth equation in table 3.

The above system of equations and inequalities has multiple solutions, including of course the correct belief and action. An agent who knows the truth ($a=\alpha$,b=ß,Σ=[0]) would choose:
(14)

$$x^* = -\frac{\alpha\,\beta}{\beta^2 + \omega}$$

An example of a class of solutions which result in incomplete learning is given by the incorrect belief $v_a=v_b=1, v_{ab}=-1$, which would be reinforced by an action $\overline{x} = 1$. The quintuple $(a,b,v_a=v_b=1,v_{ab}=-1,x=1)$ constitutes a possible incorrect limit belief and action pair if it exhibits the optimality and prediction properties. This is the case for the following values of a and b:
(15)

$$\bar{a} = \frac{\alpha + \beta + \omega}{\alpha + \beta} \qquad \bar{b} = -\frac{\omega}{\alpha + \beta}$$

Given any of the beliefs characterized in this way, the agent will find it optimal to choose x=1 even if this action does not coincide with the correct limit action x^* (which would only be the case if $\omega=-(\alpha+ß)ß$).

This example shows that incomplete learning can occur, if $\omega>0$ meaning if there is a direct cost to instrument variation. Even if $\omega=0$, incomplete learning may still be optimal. However the set of possible limit beliefs and action pairs is much smaller, since an additional property of incorrect limit beliefs can be added to the list in table 3 in this case. An optimal limit action which differs from $x^*=\alpha/ß$, only occurs for limit beliefs which imply $\overline{b}=0$.

3. Determining the Value and Extent of Optimal Experimentation

To answer the questions raised in the introduction concerning the speed of learning, the extent of optimal experimentation and the time series behavior induced by passive versus optimal learning, we need to calculate stationary optimal policies for the associated dynamic programming problem. The numerical algorithm is based on the standard proof of existence which proceeds to show that the functional equation is a contraction mapping with a unique fixed point which is the value function. The proof outlines an iterative method to calculate the value function based on the functional operator T: C[0,1] $\rightarrow$ C[0,1],

(16)

$$Tw = \underset{x}{Max} \left(R(x,\theta) + \delta \int w(B(x,\theta,\epsilon))p(\alpha,\beta \mid \theta)q(\epsilon)d\alpha \, d\beta \, d\epsilon \right)$$

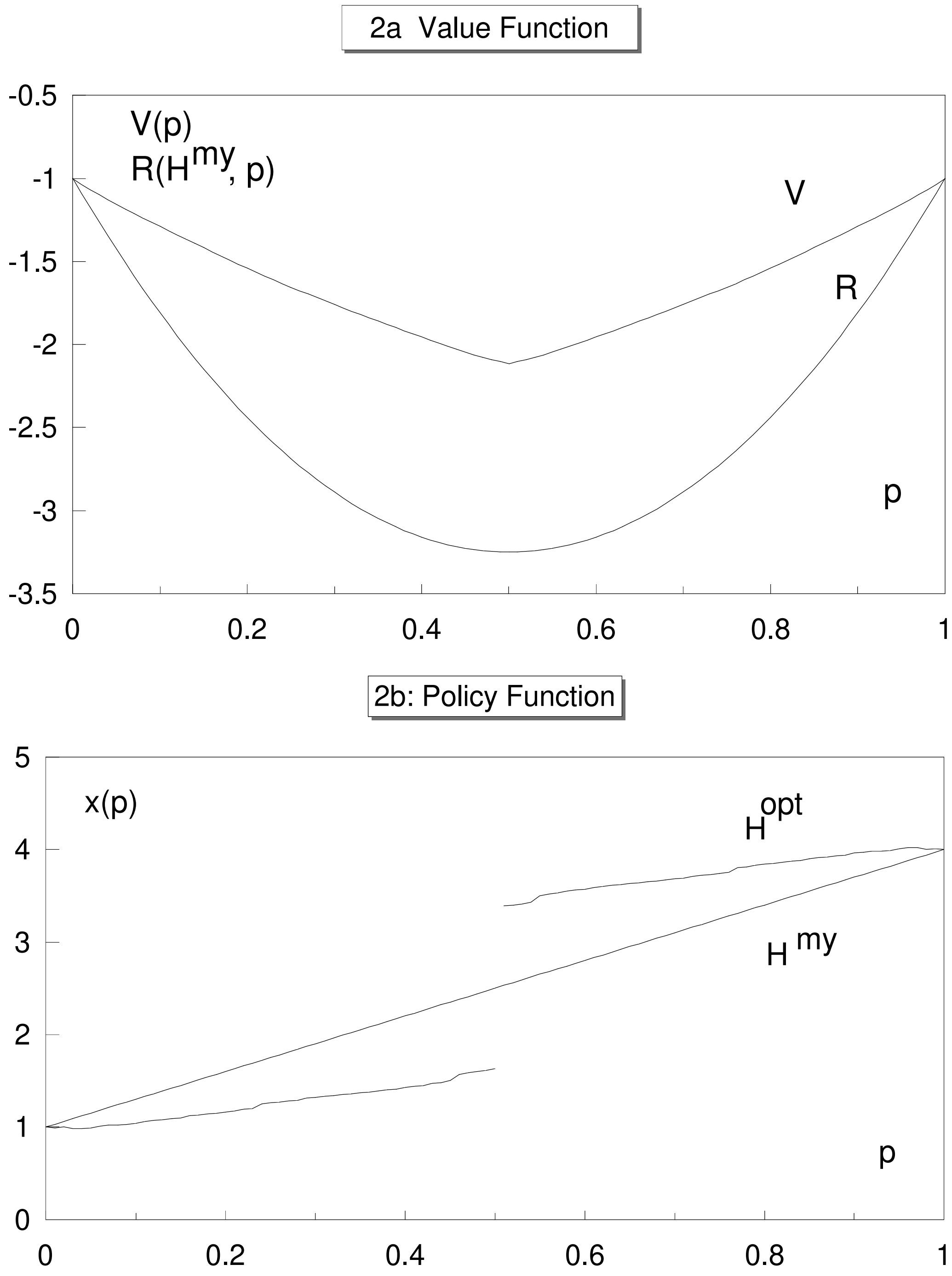
where C[0,1] is the set of bounded continuous functions. T is a contraction mapping in C[0,1] and the value function V can be calculated by direct iteration over the operator T $\{w^n \rightarrow V, n=1,2,..\}$. The necessary ingredients specified in the preceding section are: $p(\alpha,ß|\theta)$, the agent's beliefs with respect to α and ß; θ, the vector which characterizes this distribution and contains the state variables which are updated using the Bayes operator B(.); R(.), the expected one-period reward; and ϵ, the N(0,1) error term. A short summary of the methods used to implement this procedure and the associated difficulties is given in the appendices.

3.1 Illustrative Example

Approximations[12] of the value function V(p) and the optimal policy function $H^{opt}(p)$ obtained for this example with the numerical algorithm are graphed in **Figure 2**. They have been calculated on the basis of the following parameter values: $(\alpha_1,ß_1)=(-1,1)$; $(\alpha_2,ß_2)=(4,-1)$; $\delta=0.75$; $y^*=0$ and $\omega=0$. The parameter values $(\alpha_1,ß_1)$ are taken to represent the truth and consequently $(\overline{p}=1,\overline{x}=4)$ is taken to be the correct limit belief and action pair.

[12] The final discounted relative accuracy of the value function is within 0.5%. The value function displayed in figure 2a is put in `per period' terms by multiplying with $(1-\delta)$ so that it is comparable to single period reward.

Figure 2 Illustrative Example - Value and Policy Functions



(alpha1,beta1) = (4,-1) (alpha2,beta2) = (-1,1) discount factor: 0.75

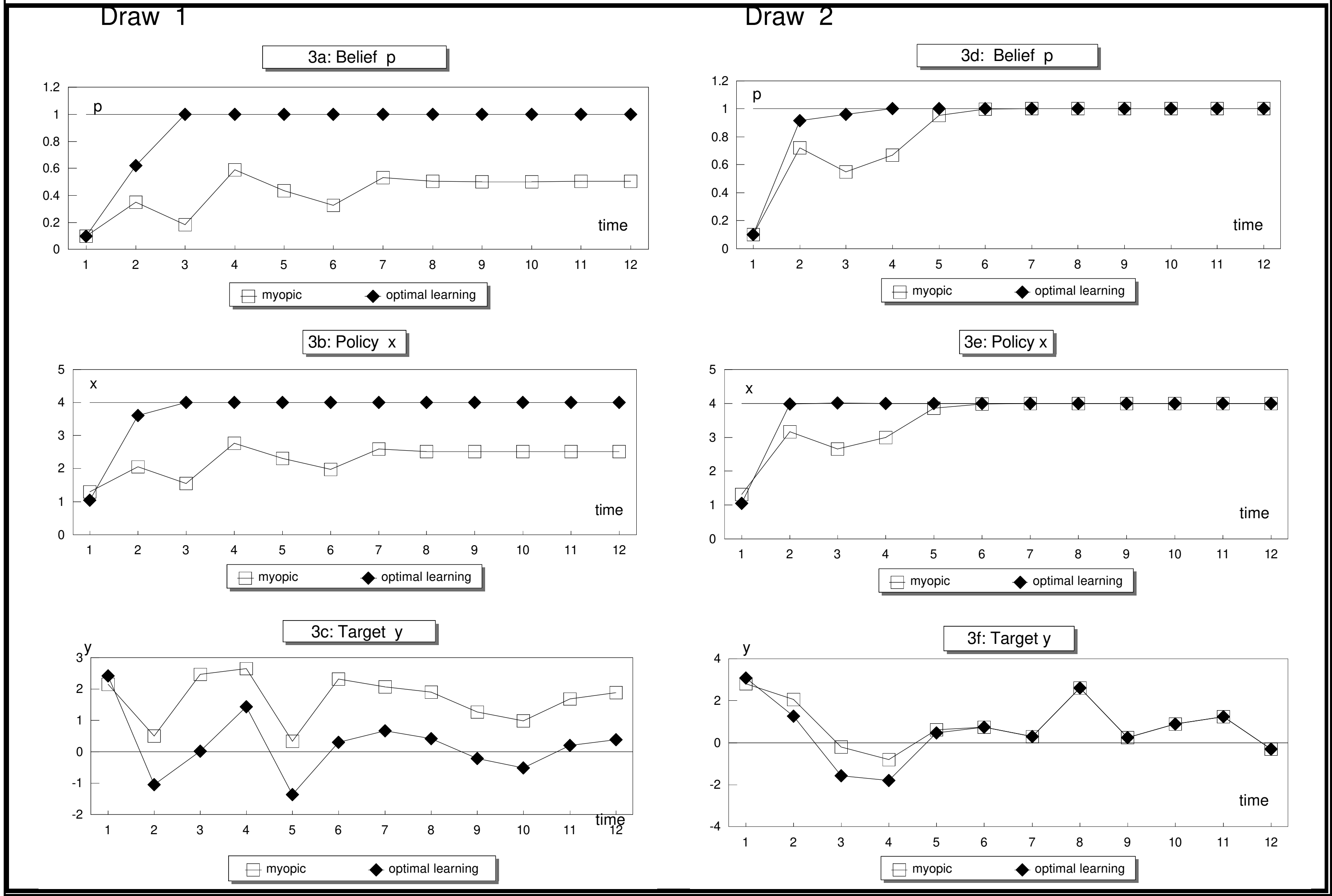
Figure 2a compares the value function V(p) to one-period expected reward $R(H^{my}(p),p)$ achieved by the one-period optimal myopic policy. While R(p) is a smooth and differentiable function, V(p) exhibits a kink at p=0.5, which means that the value function of this dynamic programming problem is not everywhere differentiable even though preferences are quadratic and the unknown data-generating process is linear. The kink point p=0.5 plays a special role in the learning process, because it constitutes the only incorrect belief which may be reinforced by a one-period optimal action, x=2.5.

Consistent with the non-differentiability in the value function, the optimal policy $H^{opt}(p)$ exhibits a discontinuity at p=0.5 as can be seen from **figure 2b**. In contrast the myopic policy $H^{my}(p)$ is a continuous and differentiable function, which averages the optimal actions under certainty (x_1=4 and x_2=1) with the weighting determined by the belief p.

The preliminary result to take away from figure 2b is that the extent of optimal experimentation, measured by the difference between passive and optimal learning policies, is large in the neighborhood of points in the state space which constitute incorrect but possibly self-reinforcing beliefs. The explanation is the following: while the passively learning/myopic agent with the belief p=0.5 would choose an action which reinforces this incorrect belief, an optimally learning agent who shares this belief would avoid this uninformative action. To the extent that the agent cares about expected future payoffs, he will accept a loss in expected current payoff in order to learn more. The larger the discount factor, i.e. the more the agent cares about the future, the closer the optimal policy will come to choosing x_1 or x_2, depending which of these actions maximizes single-period reward for the most likely point in the parameter space. For a sufficiently small discount factor, the optimal policy may not differ much from the passive learning policy and may therefore result in slow or even incomplete learning.

By simulating sample paths, we can study the speed of learning under alternative policies from a time series perspective which links the theory of optimal learning to the movement of economic observables. **Figure 3** depicts two such sample path simulations, which generate sequences of beliefs, actions and target observations for a given prior and two different draws of shocks. The chosen prior (p_0=0.1) belongs to an agent who is fairly convinced that (-1,1) is the correct set of parameters, while the true values are (4,-1).

Figure 3 Illustrative Example - Sample Path Simulations



Based on the convergence results discussed in the previous section one would expect the simulated belief and policy sequences to settle down to a fixed value, while the observations on the dependent variable should converge to a normal distribution with the mean determined by the limit action. However, whether these sequences converge to the correct or the incorrect limit belief of this example, is determined by the prior, the learning policy and the realization of shocks along each sample path.

Two observations should be made with respect to the results in Figure 3. The first draw results in complete learning under the optimal policy, but in convergence to the incorrect belief and action pair ($p=0.5, x=2.5$) under the passive learning policy. The consequence of passive learning is a persistent bias in the dependent variable y, ($|y-y^*|>0$), due to mistaken beliefs about the underlying parameters. Optimal experimentation instead would lead to complete learning at apparently little cost in terms of initial variation of x and y. For the second draw of shocks both policies result in complete learning, however a bit more slowly in the case of the passive learning policy, which biases y upwards for the first four periods.

These simulations illustrate the trade-off between current control and estimation, but the results depend on the specific prior belief and draw shocks chosen for the simulations and lack any practical relevance because of the extremely simple specification of the learning problem. Thus, to provide useful answers to the questions posed in the introduction, we need to conduct large-scale simulation exercises under a more general specification of beliefs.

3.2. Value and Extent of Optimal Experimentation: General Case

Experimentation becomes a much more important issue once we take into account that the agent cannot pin down the parameter set to two elements but rather has to assume that the unknown parameters may take any real value. The corresponding dynamic programming problem specified in table 2 models the agent's beliefs as a bivariate normal distribution $p(\alpha,\beta|a,b,\Sigma)$. In this case there exist multiple uninformative actions which may reinforce incorrect beliefs about the unknown parameters. Consequently there are multiple possible limit belief and action pairs, which are all characterized as solutions to the set of equations and inequality conditions reported before in table 3.

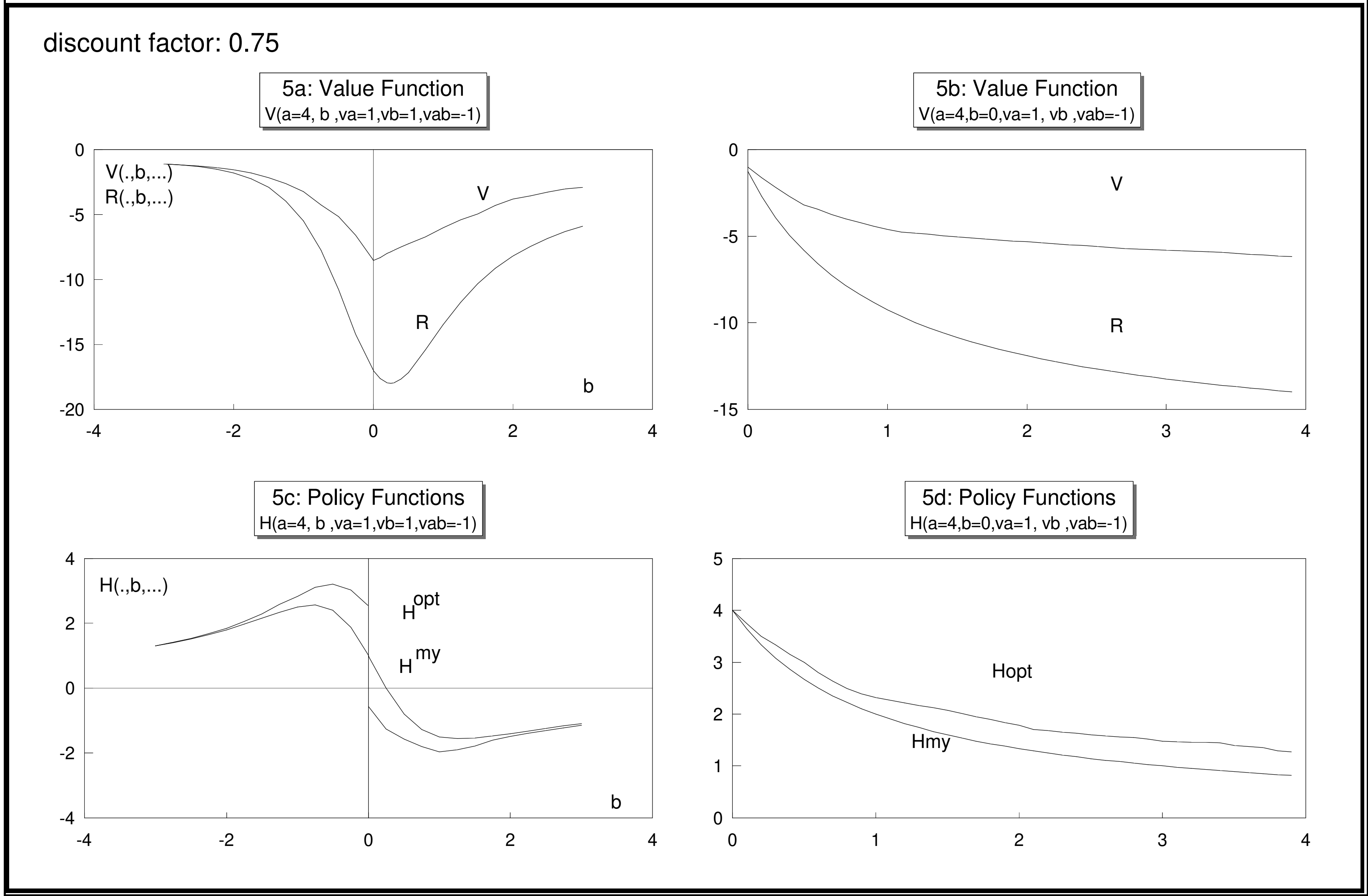
The intuition obtained from the illustrative example is that experimentation, measured by the difference between the optimal and the passive learning policies, is large in the neighborhood of points in the state space that constitute incorrect beliefs that are self-reinforcing under a one-period optimal policy. Since this general specification exhibits multiple beliefs of this kind, experimentation should be more important than in the illustrative example.

A comparison of the approximated value and policy functions $V(a,b,\Sigma)$ and $H^{opt}(a,b,\Sigma)$ to the one-period optimal policy $H^{my}(a,b,\Sigma)$ and the associated one-period reward $R(H^{my},a,b,\Sigma)$ confirms this intuition. Again we find that the set of possible incorrect limit beliefs is associated with non-differentiabilities in the value function and discontinuities in the optimal policy. Because reward, value and policies are functions of all five state variables (the means, variances and covariance), they cannot be plotted in the same straightforward manner as in the preceding example. Instead the charts in **Figure 4** plot different slices of these functions along a given dimension for fixed values of the other state variables.

The point of reference for these charts is $\hat{\theta}=(\hat{a},\hat{b},\hat{v}_a,\hat{v}_b,\hat{v}_{ab})=(4,01,1,-1)$ which represents an incorrect belief that would be reinforced by the one period optimal action, $\hat{x}=1$. Thus $(\hat{\theta},\hat{x})$ is a solution to the equations and inequality conditions of table 3 and constitutes a possible incorrect limit belief and action pair. The charts 4a and 4c, which plot slices along b, the mean of the slope parameter, show that the value function exhibits a kink and the optimal policy a discontinuity at $\hat{\theta}$. As a consequence the extent of experimentation, i.e. the difference between the myopic and optimal policies, is large in the neighborhood of $\hat{\theta}$. By definition $\hat{x}$ equals the myopic action x^{my}, which maximizes single-period expected reward based on the belief $\hat{\theta}$. Instead, the optimal action obtained from the approximated policy function is x^{opt}=2.545 which would avoid reinforcing this incorrect belief. Charts 4b and 4d provide plots for a point close to $\hat{\theta}$, which show that the value and extent of optimal experimentation remain large for other values of v_b which was kept fixed in 4a and 4c.

There are many belief and action pairs such as $(\hat{\theta},\hat{x})$, all of which satisfy the conditions summarized in table 3. For all grid-points of the value and policy functions, which are associated with such self-reinforcing incorrect beliefs, we find that the value function is non-differentiable and the optimal policy exhibits a discontinuity.

Figure 5 Two Unknown Parameters - Value and Extent of Optimal Experimentation



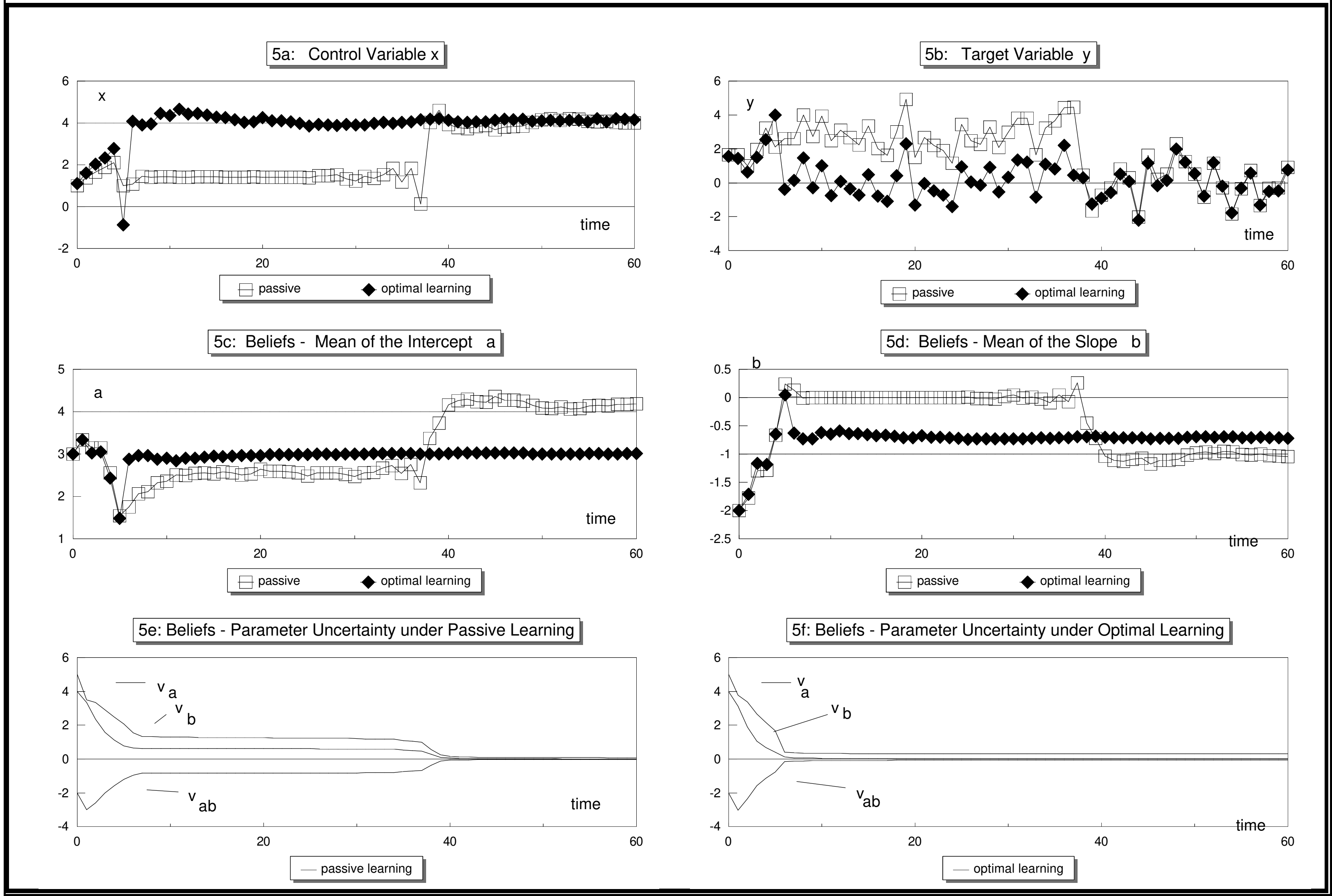
Again, we can simulate sample paths to assess the speed of learning and the frequency of convergence to incorrect limit beliefs under alternative policies. **Figure 5** compares two such simulations, one based on a myopic policy which learns passively, and the other based on an optimal learning policy. In both cases the agent starts with the same prior belief characterized by $(a_0,b_0,v_{a,0},v_{b,0},v_{ab,0})=(3,-2,5,4,-2)$ and is confronted with an identical draw of shocks. The resulting sequences of the control and the dependent variable are reported in chart 5a and 5b respectively for the case passive versus optimal learning. The sequences of the five state variables which characterize the agent's beliefs are graphed in figures 5c to 5f respectively.

Several interesting observations can be made in figure 5. First, passive learning generates a sizeable and persistent bias in the control and target sequences, as well as in the sequences of beliefs. For a long time, almost 40 periods, the agent barely changes his beliefs about the unknown parameters α and ß, even though his estimates differ strongly from the true underlying values of (4,-1). For example, during this time the means a_t, and b_t, plotted in charts 5c and 5d remain close to values of 2.4 and 0 respectively and the degree of uncertainty about these parameters as captured by the variances changes little. Due to these beliefs the agent keeps the control variable close to 1.4, which differs strongly from the correct action under certainty, $x^*=4$. As a consequence, the target observations exhibit an upward bias of about 2.6 for 40 periods.

Secondly, after this long period of time the agent is induced to take a few very informative actions, which result in rapid learning and a shift in the control and target sequences shift towards a new steady state. From the viewpoint of a researcher who does not know the data-generating process, these time series appear non-stationary, even though the underlying process is completely stationary.

Thirdly, the optimal learning policy clearly avoids such a bias, since the agent's actions converge after some initial experimentation quickly to the value which centers the target observations at $y^*=0$. Optimal experiments push the agent's beliefs away from the domain of attraction created by an incorrect belief such as $(\hat{a},\hat{b},\hat{v}_a,\hat{v}_b,\hat{v}_{ab})=(2.5833,\ 0,\ 1.204,\ 0.6,\ 0.85)$ which would be reinforced by a one-period optimal action x=1.4166. These experiments seem to incur only a small initial cost in terms of increased variability of the target variable during the first few periods.

Figure 5 Passive versus Optimal Learning about Two Unknown Parameters



Fourthly, even though at first beliefs converge rapidly towards the truth due to optimal experiments, the speed of convergence slows down dramatically after about 15 periods. The reason is that once the control sequence $\{x_t\}$ approaches the steady state value $x^*=4$ there is little remaining variance in the control variable, which constitutes the only explanatory variable of the regression.

4. Learning-by-Doing and the Time Series Behavior of Economic Observables

The sample path simulations in section 3 indicate that learning affects the behavior of economic agents in a special way and therefore has important implications for the time series behavior of economic observables generated by the actions of these agents. If learning occurs rapidly due to active decisions by the economic agent, then we should observe a higher initial variance of the control and target variables and then a rapid decline in these variances. Passive learning, which can be due to myopia or to an approach that separates control and estimation, may cause a persistent bias in the control and target sequences as long as the agent keeps choosing control values that tend to reinforce mistaken beliefs about the unknown parameters. If the agent learns the truth at a later stage, an outside observer might take the resulting shift in the control and target sequences as evidence for non-stationarities in the data even though the underlying true stochastic process is completely stationary.

Since the time series behavior observed in these simulations might be idiosyncratic to the specific draw of shocks, larger-scale simulation exercises are required to assess how frequently these properties would appear in the data.[13] Thus, I simulate 1000 sample paths with the same prior, $(a_0,b_0,v_{a,0},v_{b,0},v_{ab,0})$=(3,-2,5,4,-2), to determine how quickly actions converge on average to the correct action under certainty.

The speed of convergence as well as the extent to which a control bias emerges depends on the policy implemented, i.e. on the discount factor of the agent. An agent with a zero discount factor who does not care about future payoffs, will find it optimal to implement a myopic policy. The higher the discount factor, the more the agent will be willing to experiment, and thus reduce any bias which would be due to slow learning.

Furthermore the speed of learning is affected by the payoff parameter ω. The higher ω,

---

[13] This simulation approach goes back to Taylor (1976) and Taylor and Anderson (1976), who compare the performance of one-period optimal rules and least squares certainty equivalence rules. The latter is obtained by treating the unknown parameters as known and equal to their least squares estimates. Thus $x_t^{cert}=-a_t/b_t$ if the preference parameters y_* and ω are set to zero. They find that the certainty equivalence policy is consistent, i.e. that it converges to the policy appropriate to the true parameter values. Parameter estimates however converge extremely slowly. The conditions on limit belief/action pairs presented in table 3, can be used to confirm these results. The drawback is that certainty equivalence rules result in much larger variability during the early periods, than either the myopic or the optimal learning policy considered in this paper.

the more likely is a bias in the resulting time series, not just because it assigns a direct cost to control variability, but mainly because $\omega>0$ implies a larger set of solutions to the equations in table 3 and therefore a larger set of possible incorrect limit belief and action pairs.

I have conducted several simulation exercises which compare the time series behavior resulting from passive versus optimal learning and measure the effect of different values of δ and ω. **Table 4** reports three types of measures: **(i)** the percentage of sample paths (out of 1000) which still exhibit a control bias larger than 1 after the 20th (40th) period. The control bias is defined as (x_t-x^*) where x^* is the optimal action when the parameter values are unknown, i.e. $x^*=4$, if $\omega=0$ and $(\alpha,\beta)=(4,-1)$; **(ii)** the average control and target bias of these selected sample paths during the first 20 (40) periods, as defined by

(17)

$$\text{control bias: } \sum_{i=1}^{N}\sum_{t=1}^{T}(x_{t,i}-x^*) \qquad \text{target bias: } \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{t,i}-y^*)$$

where T equals 20 (40) and N is the number of sample paths selected. **(iii)** the average control and target bias for all 1000 sample paths.

Table 4. Passive versus Optimal Learning (ω=0)

	passive learning		optimal learning					
	(δ=0)		**(δ=0.25)**		**(δ=0.75)**		**(δ=0.95)**	
	t=20	t=40	t=20	t=40	t=20	t=40	t=20	t=40
% of Paths: Bias>1 at t	11%	2.3%	6.9%	3.5%	1.1%	0.3%	0.8%	0.4%
Biased Paths:								
Control Bias	-2.55	-2.46	-2.16	-2.16	-1.39	-1.51	-0.83	-0.81
Target Bias	2.49	2.41	1.96	2.03	1.02	1.30	0.52	0.69
All Paths:								
Control Bias	-1.14	-0.64	-0.94	-0.55	-0.56	-0.33	-0.46	-0.26
Target Bias	1.11	0.63	0.91	0.53	0.52	0.31	0.43	0.24

As shown in the upper-left corner of table 4, more than 10% of the sample paths still exhibit a sizeable bias after 20 periods, if the agent learns passively. The average control and target biases is 2.5, which is of the same size as in the simulation reported in figure 4. Even when averaging over all sample paths passive learning still results in a sizeable bias. The optimal learning policy reduces this bias and leads to faster learning the higher the discount factor. For δ=0.95 less than 1% of the sample paths still exhibit a bias after 20 periods and the average bias is reduced by more than half.

The percentage of sequences which exhibit a persistent bias increases dramatically when a direct cost is imposed on control variation (ω>0). [14] **Table 5** reports the measures described above for payoff functions with different weights ω=(0.1,0.3,0.5).

Table 5. Learning when Instrument Variation is Costly

	passive learning						**optimal learning**	
	(ω=0.1,δ=0)		**(ω=0.3,δ=0)**		**(ω=0.5,δ=0)**		**(ω=0.3,δ=0.75)**	
	t=20	t=40	t=20	t=40	t=20	t=40	t=20	t=40
% of Paths: Bias>1 at t	18%	2.3%	38%	25.4%	43.7%	33.1%	3.2%	2.7%
Biased Paths:								
Control Bias	-2.23	-2.07	-1.67	-1.58	-1.41	-1.38	-1.29	-1.23
Target Bias	2.15	2.01	1.62	1.55	1.36	1.35	1.17	1.16
All Paths:								
Control Bias	-1.17	-0.73	-1.15	-0.87	-1.00	-0.83	-0.57	-0.39
Target Bias	1.13	0.72	1.11	0.85	0.98	0.82	0.54	0.37

For higher values of ω , the bias occurs more frequently, until for ω=0.5 almost every second

[14] Payoff functions which impose such an additional cost on instrument variability are more appropriate for most applications. For example, a monopolist will incur costs in terms of inputs when varying output quantities to learn more about market demand. Similarly, a policymaker who uses the short-term interest rate as policy instrument will tend to avoid `excessive' instrument variability because it would increase volatility in financial markets.

simulated path has this property. The fact that the average target bias does not seem any larger than in table 4 where ω was set to zero, is misleading. Although the percentage of biased sequences has increased which tends to raise the average bias, the long-run target value y^* with respect to which the bias is calculated increases along with ω and thus reduces the absolute bias. The fourth column in table 5 shows that even for positive values of ω (here ω=0.3), optimal learning results in much faster convergence to the truth and tends to reduce the control and target bias .

The ex-post value of experimentation can be assessed by calculating the discounted payoffs implied by these sequences and taking averages over all sample paths. For (δ=0.95,ω=0) the average values induced by passive and optimal learning are -63.2 and -43.7 respectively. Thus, the approximated optimal learning policy results on average in an ex post improvement of 31% over the policy which separates control and estimation.

Clearly the results presented in tables 4 and 5 depend not only on the policy and the payoff parameters δ and ω, but also on the prior beliefs of the agent, the true underlying coefficient values and the variance of the shocks. Substantial work has been undertaken to assess the impact of these input parameters with the broad result that the data generated from these simulations exhibits the properties discussed above for a large range of reasonable values. More importantly, in many applications it is possible to obtain values for some of these input parameters directly from the data, while making the others the subject of sensitivity studies. For example, Wieland (1995a) formulates the learning problem of a monetary policymaker after a major structural change such as German unification, and uses pre- and post-unification estimates of the structural parameters to specify the prior beliefs of the policymaker and the true data-generating parameters.

6. Conclusions

An important contribution of this paper is to determine the value and extent of optimal experimentation for controlling a linear regression with two unknown parameters. This case is extremely useful for economic applications of learning by doing, because both the action and the parameter space, are continuous where the latter implies that from the agent's perspective the unknown parameters can take any real value. Numerical dynamic programming methods make it possible to assess the extent of optimal experimentation, which is measured by the difference between a passive learning policy that separates control and estimation and the optimal policy. This difference captures the extent to which variation is induced in the control variable purely to raise the quality of parameter estimates and improve expected future payoffs at the cost of expected current losses.

From this analysis I obtain several new results which are relevant to the theory of optimal learning, but also have important implications for time series studies of economic observables generated by agents who are learning about their economic environment. In contrast to earlier work on more restricted learning problems, specifically the seminal paper by Prescott (1972), I find that optimal experimentation can be very important in accelerating learning and can significantly improve expected future payoffs. Experimentation has a high reward in the neighborhood of incorrect beliefs which would otherwise be reinforced by a myopic or passive learning policy which is only one-period optimal. Multiple beliefs of this kind exist in any regression problem with more than one unknown parameter, which was the case studied by Prescott(1972). I also find that these beliefs are associated with potential non-differentiablilities in the value function and discontinuities in the policy function, which shows that there exists an important class of dynamic programming problems, which does not satisfy standard smoothness assumptions.

Using the policy functions obtained in this paper, one can simulate the learning process of an agent which attempts to control a linear regression with two unknown parameters. I find that passive learning frequently results in biased control and target sequences, because the agent keeps holding on to mistaken beliefs about the unknown parameters, and chooses actions which in turn justify these beliefs. If convergence to the truth occurs after some period during which beliefs have barely changed, an outside observer may interpret the resulting shift in the control and target sequences as evidence for nonstationarities even though the underlying stochastic process is completely stationary.

Optimal experimentation drastically reduces the frequency and average size of the control and target bias. An agent which cares enough about future payoffs and learns optimally, tends to avoid uninformative actions which would reinforce incorrect beliefs about the unknown parameters. Compared to passive learning the initial control and target variance is somewhat higher, but surprisingly not by much, and it declines faster as beliefs and actions converge towards their correct values.

The analysis of this paper and the numerical algorithm it uses can be applied directly to several different subjects. One example would be learning-by-doing and technological change which has been studied by Jovanovich and Nyarko (1994) and (1995) for a much simpler specification of the learning-by-doing problem, where control and estimation can be separated. Another example is the problem of a monetary policymaker who attempts to control inflation, but lacks exact knowledge about the relationship between his policy instrument and target. Based on an empirically calibrated model of monetary policy Wieland (1995a) shows that learning has important normative and positive implications for policy-making after historical episodes of structural change.

Appendix

The Numerical Algorithm

The algorithm used in this paper calculates the value function and stationary optimal policy of the learning-by-doing problem by iterating over the functional equation of the associated dynamic programming problem. It takes advantage of the contraction mapping defined by the following functional operator T, C[0,1] $\rightarrow$ C[0,1],

(A.1)

$$Tw = \underset{x}{\text{Max}} \left(R(x,\theta) + \delta \int w(B(x,\theta,\epsilon)) p(\alpha,\beta \mid \theta) p(\epsilon) d\alpha \, d\beta \, d\epsilon \right)$$

where C[0,1] is the set of bounded continuous functions. $p(\alpha,ß|\theta)$ is the probability distribution representing the agent's beliefs about the unknown coefficients α and ß. This distribution is characterized by the vector θ which contains the state variables of the DP problem. x is the control variable and ϵ is an N(0,1) error term. R(.) denotes expected current reward and B(.) represents the Bayes operator to update the beliefs.

(A.2)

$$\rho(v,w) = \underset{\theta}{\text{Supr}} \mid v(\theta) - w(\theta) \mid$$

Note that C[0,1] is a complete and separable metric space in the sup metric defined bywhere the functions v and w are elements of C[0,1]. As shown by Kiefer and Nyarko (1989) Blackwell's sufficiency conditions are satisfied and T is a contraction mapping such that

(A.3)

$$\rho(Tw,Tv) \le \delta \; \rho(w,v)$$

Thus T has a unique fixed point v, which is the value function of the learning-by-doing problem. The optimal policy corresponds to the set of x's which maximize the right-hand side of (A.1) based on the beliefs θ.

v can be calculated by directly iterating over the operator T since $T^n w \rightarrow v$ uniformly for any w in C[0,1]. A convenient starting value w^0 is the single period reward function R(.) but the iterative procedure converges for any value. If $w^{n+1} = Tw^n$, then $\rho(w^{n+1},w^n) \le \delta \; \rho(w^n,w^{n-1})$ and after iterating $\rho(w^{n+1+i},w^{n+i}) \le \delta^{1+i} \; \rho(w^n,w^{n-1})$. This implies an upper bound on the error in approximating v by w^n:

(A.4)

$$\rho(v,w^n) \le \sum \rho(w^{n+1+i},w^{n+i}) \le \frac{\delta}{1-\delta} \rho(w^n,w^{n-1})$$

This upper bound can easily be calculated since it only depends on the discount factor and the distance between the approximations obtained from the last and the preceding iteration. The time needed for convergence within a maximal error bound can be reduced significantly by introducing policy iterations.

The computational algorithm then proceeds as follows: first, calculate starting values w^0 for a grid of points in the state space Θ and save them in a table; secondly, calculate w^1 by

applying the operator T to w^0 and update said table. This second step requires calculating the maximum in x for each of the grid values of the state variables θ. For this purpose next period's expected value is calculated by evaluating the following integral:
(A.5)

$$\int w^0(B(x,\theta,\epsilon))\, p(\alpha,\beta \mid \theta)\, p(\epsilon)\, d\alpha\, d\beta\, d\epsilon$$

The functions $w^0(.)$ and B(.) are known functions and the density functions are normal. Thus the integral can be calculated using Gaussian quadrature and values of w^0 from the table, where w(.) is evaluated in between grid points by bilinear interpolation.

Given a formula for the integral the maximization problem on the right-hand side of the functional equation can be solved by standard numerical optimization procedures. However the search for the maximum turns out to be non-trivial because the FE has multiple local maxima, which results in kinks in the value function and discontinuities in the optimal policy. Therefore I use a slow but secure optimization procedure such as golden section search supplemented by a rough initial grid search.

For each value of θ, the maximum of (A1) in x gives the value of $w^1(\theta)$ used to update the table. The maximum of $|w^1(\theta)-w^0(\theta)|$ is used to calculate the upper bound of the approximation error. Finally, the whole procedure is repeated to obtain w^2 and so on until the difference between two successive approximations is sufficiently small (< 0.5%).

Computation Costs

The numerical dynamic programming problems dealt with in this paper require an immense computational effort, which is largely an effect of the so-called curse of dimensionality. Since there are 5 state variables, each approximated by N grid points, the integration and optimization procedures described above have to be carried for each gridpoint i.e. N^5 times to complete one value iteration. Specifically finding the optimum takes considerable time, because of the existence of multiple local optima.

Two steps have been taken to reduce computation time: (i) the introduction of policy iteration techniques which reduces the number of value iterations needed for convergence, and thus the number of times that the optimization procedure has to be executed; (ii) for simple payoff functions, as shown in the following section, the number of state variables can be reduced by one, which means that the integration and optimization have to be carried out only N^4 times per value iteration.

For example, the functions presented in figure 4 have been approximated for a grid of 20,47,32 and 63 gridpoints for four state variables respectively. Convergence as defined by a 0.5% maximal error was achieved after almost 2 days on a SPARC 10 (multi-processor) work station. However, the policy function used for the results in column 4 of table 5 (δ=0.75,ω=3), had to be approximated on a grid over all five state variables, which required 15 days on the same computer.

Transforming the DP Problem to Reduce the Number of State Variables

Any simplification which makes it possible to reduce the number of relevant state variables will substantially reduce the computation time required for convergence of the DP algorithm. As shown here, one can ignore one state variable when dealing with a simple payoff

function such as:
(A.6)

$$U(y) = -(y-y^*)^2$$

In this case the value function $V(a,b,\Sigma)$ is given by the supremum of
(A.7)

$$E\left[\sum_{i=t}^{\infty} \delta^i (y_i - y^*)^2 \mid a_t, b_t, \Sigma_t\right] \quad \text{where} \quad y_i = \alpha + \beta x_i + \epsilon_i$$

Proposition: The value function $V(a,b,\Sigma)$ has the following property (for any $k \neq 0$):
(A.8)

$$V(a_t, kb_t, v_{a,t}, k^2 v_{b,t}, kv_{ab,t}) = V(a_t, b_t, v_{a,t}, v_{b,t}, v_{ab,t})$$

Proof: Consider the transformed problem

$$y_t = \alpha + \hat{\beta}\hat{x}_t + \epsilon_t \quad \text{where} \quad \hat{\beta} = k\beta \quad \hat{x} = x/k$$

If the prior on (α,β) is $N(a_t,b_t,v_{a,t},v_{b,t},v_{ab,t})$ then the prior on $\hat{\beta}$ is $N((a_t,kb_t,v_{a,t},k^2v_{b,t},kv_{ab,t})$. Since these two problems are equivalent, the proposition follows.

By setting $k = v_{b,t}^{-1/2}$ and replacing it the above equation one obtains:
(A.9)

$$V(a_t, b_t v_{b,t}^{-1/2}, v_{a,t}, 1, v_{ab,t} v_{b,t}^{-1/2}) = V(a_t, b_t, v_{a,t}, v_{b,t}, v_{ab,t})$$

Thus it is sufficient to approximate the value function for a grid of four state variables. This transformation was used in this paper to obtain some of the approximation results for the value and optimal policy functions. However, it only applies if the payoff is simply a function of the signal y_t. In table 5 a more general payoff function was used, which required approximating the value function over the full grid of five state variables.

Bibliography

AGHION, P., P. BOLTON, C. HARRIS, and B. JULLIEN (1991), Optimal Learning by Experimentation, *Review of Economic Studies*, Vol. 58, pp. 621-654.

ANDERSON,T., and J.B. TAYLOR (1976), Some Experimental Results on the Statistical Properties of Least Squares Estimates in Control Problems, *Econometrica*, Vol. 44, No. 6, November 1976, pp. 1289-1302.

AOKI, M. (1989), Optimization of Stochastic Systems: Topics in Discrete-Time Dynamics, 2nd edition, Academic Press, San Diego 1989.

BERTOCCI, G. and M. SPAGAT (1993), Learning, Experimentation and Monetary Policy, *Journal of Monetary Economics*, Vol. 32, No.1, August 1993, pp. 169-178.

BALVERS, R. and T. COSIMANO (1994), Inflation Variability and Gradualist Monetary Policy, *Review of Economic Studies*, Vol. 61, pp. 721-738.

EASLEY, D. and N.M. KIEFER (1988), Controlling a Stochastic Process with Unknown Parameters, *Econometrica,* Vol. 56, 1045-1064.

EL-GAMAL, M. and R. SUNDARAM (1993), Bayesian Economists .. Bayesian Agents: An Alternative Approach to Optimal Learning, *Journal of Economic Dynamics and Control*, Vol. 17, 355-383.

FOSTER, A. and M. ROSENZWEIG (1995), Learning by Doing and Learning from Others: Human Capital and Technical Change in Agriculture, *Journal of Political Economy*, Vol. 103.

JOVANOVICH, B. and Y. NYARKO, (1994), The Bayesian Foundations of Learning by Doing, NBER Working Paper No. 4739, May 1994.

JOVANOVICH, B. and Y. NYARKO, (1995), A Bayesian Learning Model Fitted to a Variety of Empirical Learning Curves, in Bailey, M., P. Reiss, and C. Winston (eds.), *Brookings Papers on Economic Activity: Microeconomics*, 1995, 247-305.

KENDRICK, D. (1981), Stochastic Control for Economic Models, Economic Handbook Series, McGraw Hill, New York 1981.

KIEFER, N. (1989), A Value Function Arising in the Economics of Information, *Journal of Economic Dynamics and Control*, 13, 201-223.

KIEFER, N., and Y. NYARKO (1989), Optimal Control of an Unknown Linear Process with Learning, *International Economic Review*, 30, 571-586.

MCLENNAN, A. (1984), Price Dispersion and Incomplete Learning in the Long Run, *Journal of Economic Dynamics and Control* 7, 331-347.

PRESCOTT, E., (1972), The Multi-Period Control Problem under Uncertainty, *Econometrica,* Vol. 40, No.6, November 1972.

ROTHSCHILD, M. (1974), A two-armed bandit theory of market pricing, *Journal of Economic Theory,* Vol. 9, 185-202.

TAYLOR, J.B., (1976), Methods of Efficient Parameter Estimation in Control Problems, *Annals of Economic and Social Measurement,* Vol. 5, No.3, 1976.

TAYLOR, J.B., (1974), Asymptotic Properties of Multiperiod Control Rules in the Linear Regression Model, *International Economic Review*, 15, 472-484.

TREFLER, D. (1993), The Ignorant Monopolist: Optimal Learning with Endogenous Information, *International Economic Review*, Vol. 34, No. 3, August 1993.

WIELAND, V. (1995a), Monetary Policy and Structural Change: Controlling the Policy Target when Structural Parameters are Unknown, unpublished mimeo.

WIELAND, V. (1995b), A Numerical Dynamic Programming Algorithm for Optimal Learning Problems, unpublished mimeo.

ZELLNER, A., (1971), Introduction to Bayesian Inference in Econometrics, New York, Wiley, 1971.