Learning by doing and the value of optimal experimentation

Volker Wieland*
Board of Governors of the Federal Reserve System

(forthcoming in the Journal of Economic Dynamics and Control)

Abstract

Recent research on learning by doing has provided the limit properties of beliefs and actions for a class of learning problems, in which experimentation is an important aspect of optimal decision making. However, under these conditions the optimal policy cannot be derived analytically, because Bayesian learning about unknown parameters introduces a nonlinearity in the dynamic optimization problem. This paper utilizes numerical methods to characterize the optimal policy function for a learning by doing problem that is general enough for practical economic applications. The optimal policy is found to incorporate 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, these 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.

*Keywords:*Bayesian learning, optimal control with unknown parameters, learning by doing experimentation, dynamic programming
JEL Classification: C44, C60, D81, D82

---

* I would like to thank John Taylor, Kenneth Judd, Michael Horvath, Ronald McKinnon, Andrew Levin, Athanasios Orphanides, David Wilcox, Thomas Sargent and three anonymous referees for many helpful comments and suggestions. In addition, seminar participants at Stanford, the Federal Reserve Board, University of Illinois, University of Houston, Arizona State University, the 1995 European Economic Association Meeting in Prague and the 1995 World Econometric Society Meeting in Tokyo deserve thanks for stimulating discussions. All remaining errors are my own. This paper is a revised version of chapter 1 of my Ph.D. thesis at Stanford University. I would like to thank the Institute of International Economic Studies at Stockholm University for the hospitality extended during the time that I was revising this paper. The views expressed in this paper are solely my responsibility and should not be interpreted as reflecting those of the FOMC or the staff of the Federal Reserve Board.
Correspondence: Volker Wieland, Division of Monetary Affairs, Federal Reserve Board, Washington, DC 20551, U.S.A. E-mail: vwieland@frb.gov, Fax: (202) 452 2301, Phone: (202) 736 5620

1. Introduction

Economic agents rarely have complete knowledge of all the parameters that affect the payoffs that they receive, yet often they are able to learn more about these parameters from observing the outcomes of their actions. In many economic environments the agents' actions not only influence current payoffs but also provide information, which can be used to improve future payoffs.[1] Thus, agent's are faced with a tradeoff between current control and experimentation, that is, between actions that maximize expected current payoffs based on current beliefs and actions that yield lower current payoffs but superior information content.

Recent theoretical research has investigated the limit properties of beliefs and actions for a general class of learning and control problems with unknown parameters.[2] This paper fills a gap between theory and economic application by characterizing the optimal policy and value function for such a learning problem. Based on these functions one can answer a number of questions that arise in economic applications of learning by doing. The paper focuses on the following three questions:

(i) How important is experimentation?

(ii) How likely is it that the agent learns the truth— with or without experimentation?

(iii) What are the observable implications of alternative learning strategies?

I compute the value function and the optimal policy for the problem of controlling a linear stochastic regression process with two unknown parameters. In this case, the right-hand side variable is the agent's decision variable. It has an indirect effect on the agent's payoffs through the dependent variable and potentially also a direct effect. Beliefs about the unknown parameters are continuously updated according to Bayes rule. A myopic decision rule, which treats estimation and control of the stochastic process separately and disregards the potential rewards from experimentation, can easily be derived analytically. However, computing the optimal policy is not a simple task, because the learning dynamics render the optimization problem nonlinear

[1] Economic examples of such decision problems are the monopolist's profit maximization with unknown demand, [e.g. McLennan (1984), Kiefer (1989), Trefler (1993), Balvers and Cosimano (1990) and Rustichini and Wolinsky (1995)], and optimal monetary policy in transitions, [e.g. Bertocci and Spagat (1993) and Balvers and Cosimano (1994)]. Other possible applications include economic growth and technology choice and optimal investment with unknown parameters in the production function.

[2] See Easley and Kiefer (1988), Kiefer and Nyarko (1989), Nyarko (1991) and Aghion et al. (1991).

and typically preclude analytical solution. This paper instead uses numerical dynamic programming methods. Difficulties in numerical computation arise, because modeling the dynamic behavior of beliefs tend to require a large number of state variables and because the dynamic optimization problem is characterized by multiple optima.

For illustrative purposes I focus first on a learning problem with discrete parameter space, in which the decision maker is faced with only two possible sets of parameter values.[3] This specification is not very useful for economic applications but serves as a good illustrative example, because the learning dynamics can be described using a single state variable. This state variable is the probability that one of the set of parameter values is the true one. Then, I turn to the more relevant problem of controlling a linear regression where from the agent's perspective the unknown parameters may take any real value. This specification models the agent's beliefs as a bivariate normal distribution, which raises the number of state variables to five: the two means, the variances and the covariance.

The numerical approach as well as the focus on the simple regression framework render this paper very similar in spirit to early work by Prescott (1972). Prescott computed optimal learning policies 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 under very high parameter uncertainty. This result is reversed here. I detect sizeable differences between optimal and myopic policies under a non-negligible range of initial beliefs, or in other words, a sizeable extent of optimal experimentation, when there are two unknown parameters.[4]

A further contribution of this paper is to study the link between the shape of the optimal policy and the limit beliefs investigated in the literature on Bayesian learning. Easley and Kiefer (1988) and Kiefer and Nyarko (1989) showed that there exist multiple incorrect beliefs, which if reinforced by uninformative actions represent possible limit beliefs. I find that the optimal policy typically exhibits a discontinuity at such beliefs and therefore avoids uninformative actions that would render incorrect beliefs self-reinforcing. Dynamic simulations for given initial beliefs

[3] This problem is a variant of the two-armed bandit problems studied by McLennan (1984), Kiefer (1989) and El-Gamal and Sundaram (1993).

[4] Even though I only consider one additional unknown parameter, the problem is considerably more complicated, because the number of state variables is five instead of one and because there exist multiple limit beliefs and policies.

show that optimal experimentation substantially improves the speed of learning. Myopic behavior, however, frequently leads to a bias in beliefs and actions, which may persist for quite some time. These apparent nonstationarities are potentially of great importance for time series analysis of economic observables that are generated by agents who are learning about their economic environment.

The remainder of the paper is organized as follows. The next section discusses the relationship to other research on learning and control. Section 3 presents the learning problem, reviews relevant convergence results. Section 4 specifies the two cases with discrete and continuous parameter space and section 5 reports results on the value and extent of optimal experimentation. Section 6 provides dynamic simulations of the learning process under alternative policies and section 7 concludes.

2. A brief review of related approaches to learning and control

The tradeoff between current control and experimentation has been the focus of a theoretical literature on optimal Bayesian learning as well as an engineering-related literature on dual control. While the Bayesian learning literature has mostly focussed on the question whether the decision maker will learn the truth in the long run,[5] the dual control literature has been concerned with computing numerical approximations of payoffs under decision rules that involve active experimentation or probing.[6] This paper bridges a gap between the two literatures by characterizing optimal policy and value functions in a Bayesian learning framework and relating the extent of experimentation to the question of incomplete learning.

The numerical dynamic programming approach in this paper differs in several important ways from the approach taken in the dual control literature, which is discussed in detail in

[5] See for example Taylor (1974), Lai and Wei (1982), Easley and Kiefer (1988), Kiefer and Nyarko (1989), Nyarko (1991) and Aghion et al. (1991). Exceptions are Prescott (1972), Kiefer (1989) and El-Gamal and Sundaram (1993), who focus on computing optimal policies in a Bayesian learning framework.

[6] See for example Tse and Bar-Shalom (1973), Norman (1976), Norman, Norman and Palash (1979), Kendrick (1978), (1981) and (1982), Bar-Shalom (1981), Mizrach (1991), Tucci (1997), Amman and Kendrick (1994), (1995) and (1997). While the Bayesian learning literature has focused on controlled regression processes, the dual control literature has dealt exclusively with models which include dynamics in state variables other than beliefs. In this line of research active learning dual control has been found superior to passive learning in at least two different models with one control variable, either one or two state variables and either one, two or eight parameters. However, no clear superiority was found in a model with two control variables.

Kendrick (1981). First, both approaches are based on the Bellman equation associated with the dynamic learning problem, but while the dual control algorithm is formulated in a finite horizon framework without discounting, the algorithm in this paper is formulated in an infinite horizon framework with discounting. An advantage of the latter framework is that the policy and value functions are stationary. Furthermore, the algorithm is a contraction mapping and converges to the true value and policy functions. Second, the dual control algorithm approximates the optimal decision and associated payoffs by Monte Carlo simulation, while the algorithm in this paper computes expectations of future payoffs by direct numerical integration. Third, the key difficulty in computing the expected payoff for a specific value of the control variable based on the Bellman equation is that it requires an approximation of the continuation value given the state (the beliefs) in the next period. Given an approximation of the continuation value the optimal control can be found by a gradient procedure or a grid search. In the dual control literature this continuation value is typically referred to as the cost-to-go. It is approximated by a second-order Taylor expansion of the objective and updating equations about a nominal path. The algorithm in this paper instead employs an approximation of the continuation value that is based on a grid over the state space. This grid-based approximation of the value function is improved in successive iterations. By utilizing an approximation over a large range of beliefs (i.e. over a large segment of the state space) this approach may be more effective in revealing special properties of the value and policy functions such as the non-differentiabilities and discontinuities identified in this paper.[7]

The main drawback of the dynamic programming algorithm used here, is that it quickly falls victim to the "curse of dimensionality" as it is applied to models with more unknown parameters, where more state variables are needed to describe beliefs. The simulation-based dual control algorithm is less affected by this problem and can be applied to larger models.[8] Thus, the

[7] It should be noted that Kendrick (1978), and Norman et al. (1979) detected nonconvexities in the cost-to-go, which were studied in more detail by Mizrach (1991), Amman and Kendrick (1995) and Tucci (1996) and also by Balvers and Cosimano (1993) in a different framework. These nonconvexities imply the existence of multiple local optima that are consistent with my finding of non-differentiabilities in the value function and discontinuities in the optimal policy.

[8] Since the computational costs associated with a single Monte Carlo run for a given initial belief using a Taylor expansion of the cost-to-go do not increase geometrically with the number of state variables, it is feasible to conduct simulations of the dual control algorithm for models with more than two unknown parameters. However, to match

two approaches are complementary.

3. The decision problem

This paper considers the problem of a single decision maker, who attempts to control a linear stochastic process with two unknown parameters:[9]

$$y_t = \alpha + \beta x_t + \epsilon_t \tag{1}$$

First, the agent chooses a value for the control x_t given his beliefs about α and β, which are based on information prior to time t. Then, a white noise shock ϵ_t occurs and a new realization y_t can be observed. The shock ϵ_t is assumed to be normally distributed with mean zero and known variance σ^2.[10] Before choosing next period's control x_{t+1}, the agent updates his beliefs using the new information (x_t, y_t). These beliefs are modeled as a probability distribution $p(\alpha,ß|\theta_t)$, where θ_t is a vector of state variables describing the distribution.[11] A posterior distribution is obtained using Bayes rule,

$$p(\alpha,\beta \mid \theta_{t+1}) = \frac{p(y_t \mid \alpha,\beta,x_t,\theta_t)\, p(\alpha,\beta \mid \theta_t)}{p(y_t \mid x_t,\theta_t)} \tag{2}$$

which implies a set of nonlinear updating equations for the state variables denoted by

$$\theta_{t+1} = B(\theta_t,x_t,y_t) \tag{3}$$

These updating equations represent the learning channel, through which the current action x_t affects next period's beliefs and thus indirectly future realizations of x and y. Both, the control x and the signal y, may affect the agent's payoffs as defined by the objective function $U(y,x)$. In this paper I consider a quadratic objective function

---

the results in this paper, it would be necessary to conduct runs for all initial beliefs that constitute grid points in the grid-based approximation. Since there are about $2x10^6$ such grid points, I would expect that computational costs using dual control methods would also be substantial, but a more direct comparison would be of interest.

[9] This framework can be extended to include lagged dependent variables which would introduce additional dynamics. I plan to consider this case in future research. The advantage of the simple regression is that all the dynamics are due to learning and only state variables describing the agent's beliefs need to be considered.

[10] This assumption is convenient, because it implies that posterior beliefs in the specifications studied hereafter are normal distributions.

[11] For example, in the case of bivariate normal beliefs this vector contains the means, variances and covariance.

(4)

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

with a desired target level y^* and a weight $\omega \geq 0$. For $\omega=0$ this coincides with the input-target model that is often used in studies of learning by doing [e.g. Jovanovich and Nyarko (1996) and Foster and Rosenzweig (1995)]. However, the numerical algorithm developed in this paper admits other functional forms and could be adapted to specific economic applications. Taking expectations with respect to α,ß and ϵ one obtains expected one-period reward $R(x,\theta)$,

(5)

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

where $q(\epsilon)$ stands for the normal density function of the shocks. A decision maker who treats control and estimation separately, will choose x to maximize $R(x,\theta)$ based on current parameter estimates. After observing y, he will proceed by updating the parameter estimates and selecting next period's control. This behavior is myopic since it disregards the effect of current actions on future beliefs.

A stationary policy or feedback rule is defined as a function $H(\theta)$, which selects an action x based on the current state θ. Given a specification of the agent's beliefs, the myopic policy $H^{my}(\theta)$, which maximizes $R(x,\theta)$, can be derived analytically. This is not the case for the optimal policy $H^{opt}(\theta)$, which maximizes the discounted sum of expected current and future rewards

(6)

$$\begin{aligned} &\text{Max } E\left[\sum_{j=0}^{\infty} \delta^j R(x_{t+j}, \theta_{t+j}) \mid \theta_t, H\right] \\ &[x_{t+j}]_{j=0}^{\infty} \\ &\text{s.t. (1) , (3)} \end{aligned}$$

where δ denotes the discount factor. The expectation is with respect to future beliefs (θ_{t+j}, j=1,2,...) and is conditional on the initial prior belief θ_t as well as the policy $H(\theta)$. If α and ß were known, this optimization problem would reduce to a static problem and the myopic policy would be optimal. However with α and 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 (θ_{t+j}, j=1,2,...) depend on the entire sequence of actions up to that point (x_{t+j-1},j=1,2,...) and thus on the function $H(\theta)$. The effect of policy on future beliefs and the expectations operator is also apparent from the Bellman equation associated with this dynamic programming problem: (7)

$$V(\theta) = \underset{x}{\text{Max}} \left[R(x,\theta) + \delta \int_{\mathbb{R}} \int_{\mathbb{R}^2} V(B(\theta, x, \alpha + \beta x + \epsilon))\, p(\alpha, \beta \mid \theta)\, q(\epsilon)\, d\alpha\, d\beta\, d\epsilon \right]$$

$V(\theta)$ denotes the value function and the two terms on the right-hand side characterize the tradeoff between current control and estimation. The first term is current expected reward, while the second term is the expected continuation value in the next period, which reflects the expected improvement in future payoffs due to better information about the unknown parameters. Next period's beliefs have been substituted out of next period's value function using the Bayes operator $B(.)$. They depend on the realization of the dependent variable $y=\alpha+ßx+\epsilon$ whose conditional distribution is a function of x, the beliefs $p(.)$ and the distribution of the error term $q(.)$.

As shown by Easley and Kiefer (1988) and Kiefer and Nyarko (1989) a stationary optimal policy exists and the value function is continuous and satisfies the Bellman equation.[12] Policy and value functions can be obtained using an iterative algorithm based on the Bellman equation starting with an initial guess about $V(.)$. However, the integration in (7) can usually not be carried out analytically, because $B(.)$ is a nonlinear function of y and x. Thus, there are many examples, including the cases considered in this paper, for which no analytic solutions have been found even though the unknown stochastic process is linear and the payoff function is quadratic. The feasibility of numerical approximation depends on the specification of the agent's beliefs.

It remains to discuss the asymptotic properties of beliefs and actions, which have been the main focus of Easley and Kiefer (1988) and Kiefer and Nyarko (1989) (KN hereafter). Standard convergence results are not applicable, because along any sample path for which parameter estimates converge, the sequence of actions also converges. If actions converge too

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

rapidly they may not generate enough information for identifying the unknown parameters and the limit distribution representing the agent's limit belief need not be centered on the true parameter values. KN show that the process of posterior beliefs converges to a limit belief $\overline{\theta}$ for any multiple linear regression process under minimal distributional assumptions.[13] However, this convergence result does not pin down the limit belief itself. There may exist multiple limit beliefs that are outcomes of optimal policy but do not coincide with the true parameter values. Incorrect beliefs may be self-reinforcing, because learning is costly and actions that would be sub-optimal under the truth, may be optimal under these subjective incorrect beliefs.

For the case of a simple regression with known error distribution, KN show that all limit belief and policy pairs $(\overline{\theta}, \overline{x})$, whether incorrect or not, share three properties, *belief invariance*, *one-period optimization* and *mean prediction*, which can be used to describe the set of possible limit beliefs. (8)

$$\textbf{Belief Invariance:} \quad \overline{\theta} = B(\overline{\theta}, \overline{x}, \alpha + \beta\overline{x} + \epsilon)$$

$$\textbf{One-Period Optimization:} \quad R(\overline{x}, \overline{\theta}) = \text{Max}_{x} E\left[U(\alpha + \beta x + \epsilon, x) \mid \overline{\theta} \right]$$

$$\textbf{Mean Prediction:} \quad E[\alpha \mid \overline{\theta}] + E[\beta \mid \overline{\theta}]\,\overline{x} = \alpha + \beta\,\overline{x}$$

First, *belief invariance* simply follows from the convergence result. For a belief to be a limit belief it needs to self-reinforcing, that is, given the limit action $\overline{x}$, updating according to Bayes rule should again generate the limit belief $\overline{\theta}$. Thus, limit belief and policy pairs define fixed points of the updating equations $B(.)$. Second, with invariant beliefs the dynamic optimization problem reduces to the static problem of maximizing expected one-period reward. Thus, *one-period optimization* refers to the fact that the limit action $\overline{x}$ maximizes expected one-period reward conditional on the limit belief $\overline{\theta}$. Third, if the control variable is held constant at $\overline{x}$

[13] The distributional assumptions made here are useful in computing the optimal policy. KN prove convergence of the posteriors by means of the martingale convergence theorem without restricting beliefs to any conjugate families. The process of posterior beliefs is a martingale, which implies that the agent does not expect his beliefs to change in any predictable manner. It is straightforward to show that this property applies to the probability p in the illustrative example as well as to the means in the bivariate normal specification.

forever, the agent will at a minimum learn the associated mean value of the dependent variable *y* in the limit, which constitutes the *mean prediction property* of limit beliefs.

4. Two specifications of the agent's beliefs and the associated limit properties

Hereafter, each result will first be discussed for the illustrative example with discrete parameter space, which avoids the complexity that would arise from multiple state variables and multiple limit beliefs, and then for the more general specification with continuous parameter space.

4.1 Illustrative example with discrete parameter space

There are two possible sets of parameter values $(\alpha_1,ß_1)$ and $(\alpha_2,ß_2)$. Consequently the vector θ_t which characterizes the agent's belief has a single element—the probability p that $(\alpha,ß)=(\alpha_1,ß_1)$. Table 1 specifies the updating equations, expected current reward, the myopic policy and the associated Bellman equation.

TABLE 1 ABOUT HERE

The objective is defined by the quadratic payoff function as in equation (4) with $y^*=0$ and $\omega=0$ and the variance of shocks σ^2 is set equal to one. The optimal policy under certainty is independent of p and is equal to $x_1=-\alpha_1(ß_1)^{-1}$ or $x_2=-\alpha_2(ß_2)^{-1}$, whichever are the true parameter values. If $(\alpha_1,ß_1)$ are the true values, the correct belief is $p=1$, otherwise it is $p=0$. The myopic policy is simply a weighted average of the optimal actions under certainty, x_1 and x_2, with the weight depending on the belief p. Learning the true parameter values implies that the belief p converges either to one or to zero, while the policy converges either to x_1 or to x_2.

The three properties of limit belief and policy pairs discussed in section 2, belief invariance, one-period optimization and mean prediction, can be used to show that there exists at most one incorrect limit belief. First, belief invariance implies that limit belief and action pairs are fixed points of the updating equation. By definition $(\overline{p}=0, \overline{x}=x_2)$ and $(\overline{p}=1, \overline{x}=x_1)$ are such fixed points, because the truth is a possible limit belief. Furthermore, any $\overline{p}$ associated with an action $\overline{x}$ s.t. (9)

$$\alpha_1 + \beta_1 \overline{x} = \alpha_2 + \beta_2 \overline{x}$$

constitutes such a fixed point. This belief and action pair also exhibits the mean prediction property. Finally, it follows from one-period optimization that (10)

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

The solution to the two-equation system (9) and (10) for $\bar{p}$ defines the only possible incorrect limit belief: (11)

$$\bar{p} = \left(1 - \frac{\beta_1}{\beta_2}\right)^{-1}$$

Note that $\bar{p}$ is a well-defined probability, $\bar{p}\in[0,1]$, only if the slope parameters ß$_1$ and ß$_2$ are of opposite sign.

FIGURE 1 ABOUT HERE

Figure 1 provides some further intuition. The two curves are based on parameter values of (α_1=4,ß$_1$=-1) and (α_2=-1,ß$_2$=1) which imply that x_1=4 and x_2=1. From (10) and (11) it follows that the incorrect limit belief and action pair is ($\bar{p}$=0.5, $\bar{x}$=2.5). Since $\bar{x}$=2.5 corresponds to the intersection of the two curves, this value does not provide any information that would allow the decision maker to distinguish between the two curves. Such an uninformative action would of course reinforce any belief held by the decision maker. For this action to be a limit action, it needs to maximize expected one-period reward given some belief. As is apparent from (10) any one-period optimal action is associated with a value of p between 0 and 1 and lies on the interval [x_2=1, x_1=4]. Thus, as long as the intersection point lies on this interval, there exists one incorrect limit belief and action pair. Note that if the two curves had slopes of the same sign, the intersection point would fall outside of the interval. In this case, the uninformative action would not be one-period optimal and the only possible limit belief would be the truth.

4.2. General specification with continuous parameter space

A specification with continuous parameter space promises to be more useful for studying

the implications of learning by doing in economic models, because implicit is the more realistic assumption that from the agent's perspective the unknown parameters may take any real value. For this purpose I specify the agent's beliefs as a bivariate normal distribution: (12)

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

The state vector θ_t contains five variables, the means (a_t,b_t), the variances $(v_{a,t},v_{b,t})$ and the covariance $v_{ab,t}$. Σ_t denotes the variance-covariance matrix. Table 2 shows the nonlinear updating equations,[14] the expected current reward for a quadratic objective, the myopic policy and the Bellman equation under bivariate normal beliefs.

TABLE 2 ABOUT HERE

Using the three properties of limit beliefs, belief invariance, mean prediction and one-period optimization, one can show that there exist multiple limit beliefs—all but one incorrect. The correct limit belief is $(\overline{a}=\alpha, \overline{b}=ß, \overline{v}_a=\overline{v}_b=\overline{v}_{ab}=0)$ and implies the following choice of the control:

(13)

$$x^* = -\frac{(\alpha - y^*)\beta}{\beta^2 + \omega}$$

The set of possible limit beliefs and policies $(\overline{a},\overline{b},\overline{v}_a,\overline{v}_b,\overline{v}_{ab},\overline{x})$ is characterized by the system of four equations and three inequality conditions in table 3.

TABLE 3 ABOUT HERE

This system of equations and inequalities has multiple solutions. Further manipulation leads to a more concise description of the set of incorrect limit beliefs. It follows from belief invariance that incorrect limit beliefs are characterized by uncertainty and perfect correlation (a correlation coefficient ρ of unity) and that the limit action equals the negative ratio of $\overline{v}_{ab}$ and $\overline{v}_b$.

[14] A derivation of the updating equations using Bayes rule can be found in Zellner (1971). In the case considered here the updating equations correspond exactly to recursive least squares or the Kalman filter. The updating is conditional on the known variance of shocks. This assumption is standard and ensures that the posterior belief is a normal distribution and the optimal policy can be computed. The updating equations here are written for a variance of shocks equal to one.

Furthermore, one-period optimization implies a relationship between this ratio and the means of the limit distribution: (14)

$$\text{Perfect Correlation: } \frac{\bar{v}_{ab}^2}{\bar{v}_a \bar{v}_b} = 1 = \rho^2$$

$$\text{Uncertainty: } \bar{v}_a, \bar{v}_b > 0$$

$$\text{Limit Actions: } \bar{x} = -\frac{\bar{v}_{ab}}{\bar{v}_b} = -\frac{\bar{v}_{ab} + (\bar{a} - y^*)\bar{b}}{\bar{v}_b + \bar{b}^2 + \omega}.$$

It is important to note that this information is available to the decision maker. Even though he does not know the true parameter values, he can derive a set of candidate incorrect limit beliefs based on (14). Using the mean prediction property and our knowledge of the true parameter values, we can also derive the means associated with the limit distribution. These means equal (15)

$$\bar{a} = \alpha + (\bar{b} - \beta)\frac{\bar{v}_{ab}}{\bar{v}_b} \qquad \bar{b} = \frac{\omega \frac{\bar{v}_{ab}}{\bar{v}_b}}{\alpha - y^* - \beta \frac{\bar{v}_{ab}}{\bar{v}_b}}$$

if the preference parameter ω is greater than zero. Thus, for any non-zero value of $\bar{v}_{ab}$ and any positive value of $\bar{v}_b$ one can use (14) and (15) to find values of $\bar{a}$, $\bar{b}, \bar{v}_a$ and $\bar{x}$ which together constitute a limit belief and action pair. Furthermore, it is now straightforward to provide examples of such incorrect limit belief and action pairs. One simply chooses values for the covariance and variance such that the negative ratio is different from the correct action given by (13), and then computes the associated limit means from (15). Such an example would be ($\bar{a}=\alpha-\beta+\omega(\alpha-y^*-\beta)^{-1}$, $\bar{b}=\omega(\alpha-y^*-\beta)^{-1}$, $\bar{v}_a=\bar{v}_b=1, \bar{v}_{ab}=-1$, $\bar{x}=1$).

Finally, if the preference parameter ω is equal to zero, it follows from the third equation in (14) that the mean of the slope parameter $\bar{b}$ will either be equal to $(\bar{a}-y^*)\bar{v}_{ab}\bar{v}_b^{-1}$ or equal to

zero in the limit. In the first case the associated limit action is $\overline{x} = -(\overline{a}-y^*)\overline{b}^{-1}$. It follows from the mean prediction property that $\overline{x} = -(\alpha-y^*)\beta^{-1}$, which is the optimal action under certainty. In other words, this limit belief, even if it is incorrect, will induce the correct action. Thus, one can conclude that if $\omega=0$, then all incorrect limit belief and action pairs will imply $\overline{b}=0$.

5. Determining the value and extent of optimal experimentation

This section addresses the questions raised in the introduction concerning the value and optimal extent of experimentation based on numerical approximations of the optimal policy and value functions. These approximations are obtained using an algorithm that is based on the Bellman equation. I describe the algorithm, its numerical implementation, the precision of the approximations and the computational costs in detail in the appendix.

5.1 Illustrative example with discrete parameter space

Figure 2 compares the optimal policy and value functions to the myopic policy and the resulting one-period and multi-period expected reward for the problem with discrete parameter space. The two sets of parameter values are $(\alpha_1,ß_1)=(4,-1)$ and $(\alpha_2,ß_2)=(-1,1)$ as in figure 1. The values of the preference parameters are $\delta=0.75$, $y^*=0$ and $\omega=0$ and the variance of the normally distributed zero-mean shocks is set to one.

FIGURE 2 ABOUT HERE

The upper panel compares the value function $V(p)$ (dotted line with each dot representing a grid point) to the expected one-period reward $R(H^{my}(p),p)$ (solid line) and the expected (infinite horizon) multi-period reward (dashed line) under the myopic policy. Both, the value function and the multi-period reward from the myopic policy are put in per-period terms by multiplying with $(1-\delta)$. The difference between them represents the expected gain from optimal experimentation, which lies between 0 and 35 percent of the multi-period reward under myopic behavior depending on the initial belief. The gain is largest for a belief of $p=0.5$. For this belief the expected multi-period reward from the myopic policy in per-period terms is equal to the one-period reward. In other words, under this belief the payoffs obtained from the myopic policy are expected to remain unchanged for all future periods. Another interesting point to notice is that $V(p)$ exhibits a non-differentiability at $p=0.5$, while $R(p)$ is smooth and differentiable. As shown in the preceding

section, p=0.5 is the single incorrect limit belief and x=2.5 is the one-period optimal uninformative action, which renders this belief self-reinforcing.

The lower panel shows that the optimal policy $H^{opt}(p)$ exhibits a discontinuity at p=0.5, while the myopic policy $H^{my}(p)$ is continuous and differentiable and coincides with the uninformative action x=2.5 for a belief of p=0.5. The discontinuity of the optimal policy has an intuitive explanation. By avoiding uninformative actions that would reinforce an incorrect belief such as x=2.5 and p=0.5 the optimal policy guarantees that the agent will eventually learn the true parameter values. The myopic policy instead would render p=0.5 self-reinforcing and allow beliefs to converge to it in the limit. The optimal extent of experimentation as measured by the difference between the myopic and optimal policy is larger, the closer the current belief is to the candidate incorrect limit belief. Numerical approximations for different values of the discount factor δ show that the discontinuity arises even with values as low as 0.25 and that the optimal degree of experimentation increases with the discount factor. The more the decision maker cares about future performance, the more he will experiment. As a consequence his decisions exhibit a tendency towards the extreme actions, x_1=4 or x_2 =1, depending on which of these actions maximizes one-period reward for the most likely point in the parameter space.

5.2. General case with continuous parameter space

Based on the above result, one might expect that the gains from experimentation are even larger in the case with bivariate normal beliefs, simply because there exist multiple beliefs that motivate uninformative actions under myopic behavior and consequently represent possible limit belief and policy pairs. Furthermore, one might expect that the value and policy functions respectively exhibit non-differentiabilities and discontinuities at such beliefs. These conjectures are confirmed by the numerical analysis.

Figure 3 compares the value function (dotted line with each dot representing a grid point) to the expected one-period (solid line) and multi-period reward (dashed line) under the myopic policy in per-period terms.

FIGURE 3 ABOUT HERE

Since these are five-dimensional functions, the nine panels in figure 3 only show slices along one dimension of the state, the expected value of the slope parameter b, for alternative values of the

other state variables. All panels are associated with an expected value of the intercept a=4 and an associated variance v_a=1. Each row of panels is associated with a different value of the variance of the slope. In the top row v_b=2.2, in the middle row v_b=1 and in the bottom row v_b=0.5. Each column is associated with a different value of the correlation coefficient $\rho=(v_{ab})(v_a v_b)^{-0.5}$. In the left column, ρ=-1, in the middle column ρ=-0.5 and in the right column ρ=0. The value of the covariance is different in each panel and can be derived from the correlation coefficient. Finally, the values of the preference parameters are again δ=0.75, y^*=0 and ω=0 and the variance of the normally distributed zero-mean shocks is set to one.

The expected gain from optimal experimentation corresponds to the difference between the value function and the expected multi-period reward from the myopic policy. It varies between 0 and 52 percent, depending on the agent's beliefs about the unknown parameters, and is typically largest for a belief of b=0. Furthermore, at b=0 expected multi-period reward from the myopic policy in per-period terms equals expected one-period reward, which means that future payoffs from the myopic policy are expected to remain unchanged. At b=0 the value function typically exhibits a kink or non-differentiability. Finally, a comparison between the different rows of panels reveals that the range of beliefs about the slope b, for which the gain from experimentation is large, increases with the variance of the slope v_b.

Figure 4 shows the myopic and optimal policies corresponding to the value and reward functions in figure 3. In all cases, the optimal policy implies a greater or equal value of the control variable than the myopic policy. The extent of experimentation is largest in the neighborhood of b=0 and the optimal policy typically exhibits a discontinuity at this point. Furthermore, a comparison of the three rows of panels confirms that the range of beliefs about the slope, for which the extent of experimentation is quantitatively significant, also increases with uncertainty about the slope.

FIGURE 4 ABOUT HERE

The non-differentiability of the value function and the discontinuity of the optimal policy at beliefs with b=0 arises because the myopic action under these beliefs does not provide valuable information. First, in the case of the beliefs underlying the three panels in the left column of figures 3 and 4, one can use (14) to show that the point b=0 represents a candidate incorrect limit belief, i.e. a belief that is self-reinforcing under myopic behavior. The correlation coefficient is

equal to -1, while the myopic policy equals the ratio of v_{ab} and v_b and is uninformative. The discontinuity of the optimal policy at this point implies that the optimal action is informative and avoids reinforcing the incorrect belief. Such a discontinuity also arises in the panels of the middle and right-hand side columns, even though the correlation coefficient is greater than -1 and b=0 consequently not a candidate limit belief. However, using the conditions in table 3 one can show that for b=0 both, the myopic policy and the resulting payoffs are expected to remain invariant under the myopic action. At b=0 the myopic action equals the ratio of v_{ab} and v_b, which in turn implies that b, v_{ab} and v_b and thus the myopic action itself are invariant. While the myopic action may provide information regarding a and $v_{a,}$ this information is not expected to improve future payoffs. Thus, the discontinuities in the optimal policy arise at beliefs, under which the myopic policy is either completely uninformative or does not provide any valuable information.

Figures 3 and 4 illustrate a more general numerical result concerning convergence under optimal behavior. As shown in section 3 there are many candidate limit belief and action pairs, which satisfy the two equations and inequality conditions summarized by (14). I find that all grid-points of the value and policy function approximations, which represent such a candidate incorrect limit belief, are associated with a discontinuity in the optimal policy. This is as close as one can get to a proof of complete learning under the optimal policy by means of numerical methods. Furthermore, I found this result to be robust to alternative assumptions concerning the variance of shocks and the preferences parameters. For example, I have tried discount factors δ between 0.25 and 0.95 and values of the weight ω between 0 and 1, and have found discontinuities in the optimal policy in all cases.

6. Learning by doing and the time series behavior of economic observables

Typically we cannot observe the beliefs of economic agents. Thus, from an empirical perspective it is of interest to investigate whether alternative assumptions concerning the learning behavior of economic agents would result in observable differences in economic time series data that are affected by the actions of these agents. One can shed some light on this question by means of dynamic simulations of the myopic and optimal policies discussed in the preceding section.

6.1 Illustrative example with discrete parameter space

Figure 5 shows the outcomes of four different dynamic simulations for the case with discrete parameters. The true parameter values in these simulations are (α_1=4, ß$_1$=-1). Thus, the correct limit belief and action pair is ($\overline{p}$=1, $\overline{x}$=4). In all four simulations, the initial prior belief is the same and implies that the decision maker assigns only a 10 percent probability to the true parameter values, i.e. p_0=0.1. Differences arise because I consider two different draws of shocks and simulate the agent's learning process for each draw under myopic and optimal behavior.

FIGURE 5 ABOUT HERE

As can be seen from the panels in figure 5, which show the simulation paths for beliefs *p*, actions *x* and observations *y*, the optimal policy (dotted line) always results in complete learning. Given the discontinuity of the optimal policy at the candidate incorrect limit belief *p*=0.5, which was discussed in the preceding section, this should come at no surprise. Convergence to the truth occurs very rapidly, essentially within two periods for both sets of shocks. As a consequence, the control variable x converges to the optimal value under certainty and the dependent variable converges to a normal distribution with the mean equal to the target value of zero and the variance equal to the variance of shocks.

Myopic behavior (dashed line) may result in convergence to the incorrect limit belief and action pair, (*p*=0.5, *x*=2.5), as in the case of the first draw of shocks. The second simulation shows that for another sequence of shocks myopic behavior may result in convergence to the correct belief for the same initial conditions. However, convergence under the myopic policy is still slower than under the optimal policy. Myopic behavior leads to a bias in the dependent variable relative to the target value y^*, which may be temporary or permanent and is due to mistaken beliefs about the underlying parameters. The next step is to explore whether such a bias may also occur in a more realistic learning problem, and if so, how frequently it arises and how persistent it may be.

6.2. General case with continuous parameter space

Figure 6 compares two dynamic simulations for the specification with continuous parameter space, one based on myopic and the other on optimal behavior. In both cases, the decision maker has the same prior belief, (a_0,b_0,$v_{a,0}$,$v_{b,0}$,$v_{ab,0}$)=(3,-2,5,4,-2), and is confronted with

the same sequence of unexpected shocks.

FIGURE 6 ABOUT HERE

As can be seen from the two panels in the top row, myopic behavior generates a sizeable and quite persistent bias in the control and target variable. The source of this bias are incorrect beliefs about the unknown parameters α and ß. The agent's point estimates of approximately (2.4,0) differ substantially from the true underlying values of (4,-1). As a consequence, the agent keeps x close to 1.4, while the appropriate decision given the true parameter values would be to set x equal to 4. The result is an upward bias in the dependent variable y of about 2.6. For more than 30 periods, the agent's myopic decisions induce almost no changes in beliefs and the degree of uncertainty remains fairly constant. This is due to the domain of attraction created by an incorrect belief such as $(a,b,v_a,v_b,v_{ab})=(2.5833, 0, 1.204, 0.6, 0.85)$, which is self-reinforcing under the myopic decision x=1.4166. While decisions and beliefs do not completely converge to this possible limit belief and action pair, they stay close to it for some time. Then some shocks occur, which induce the agent to take more informative decisions that reduce uncertainty and lead to new point estimates of the unknown parameters. As a result, the control and the dependent variables eventually converge to a steady state that coincides with the optimal outcome under certainty.

No such bias arises under optimal behavior. After some initial experimentation the agent's actions rapidly converge to x=4, and the dependent variable, on average, equals the agent's target of zero. The cost of these experiments in terms of increased variability of the dependent variable during the first few periods seems relatively small. At first, parameter uncertainty declines rapidly and the point estimates move quickly towards the true value. However, once decisions are close to what would be optimal under the true parameter values, beliefs only continue to converge very slowly towards the truth.

These simulations indicate some implications of learning, which would allow a researcher who observes the actions of economic agents but not their beliefs, to draw some inferences concerning the agents' learning processes. For example, a rapid decline in the variability of the dependent variable would tend to be associated with optimal experimentation, while apparent non-stationarities and shifts in the data would tend to be associated with myopic behavior. However, the simulation results shown in figure 6 could be idiosyncratic to the specific sequence of shocks, and a larger set of simulations is needed to assess how frequently these features would appear in

the data.[15]

In the following I report results from several such simulation exercises, each based on 1000 draws of shocks of 100 periods length and the same initial belief, $(a_0,b_0,v_{a,0},v_{b,0},v_{ab,0})$= (3,-2,5,4,-2). These exercises include simulations based on alternative values for the agent's discount factor δ and the weight ω in the loss function. The results are summarized using three types of measures: *(i)* the percentage of simulations which exhibit a control bias larger than one in absolute value after the 20th (40th) period, where the control bias is defined as (x_t-x^*) and x^* is the optimal action when the true parameter values are known, i.e. $x^*=4$, if $\omega=0$ and $(\alpha,ß)=(4,-1)$; *(ii)* the average control and target bias of this subset of simulations during the first 20 (40) periods, (16)

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

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

TABLE 4 ABOUT HERE

As shown in the first row of table 4, under myopic behavior more than 10% of the sample paths still exhibit a sizeable bias after 20 periods. The average control and target biases are -2.5 and 2.5 respectively as shown in the second and third row. Thus, conditional on the emergence of a bias, the size of the bias in the simulation reported in figure 6 is quite representative. Even when averaging over all simulations, myopic decisions still result in a sizeable bias. Under optimal learning, however, it is less likely that a bias due to incorrect beliefs about the underlying

[15] This simulation approach goes back to Taylor (1976) and Anderson and Taylor (1976), who compare the performance of one-period optimal rules and least squares certainty equivalence rules. The latter are 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-equivalent policy is consistent, i.e. that it converges to the decision that is appropriate under the true parameter values. Parameter estimates however converge extremely slowly. The conditions on limit belief and action pairs presented in table 3, can be used to confirm these results. The drawback of certainty-equivalent decision rules is that they lead to much more variability initially than either the myopic or the optimal policy considered in this paper and are always suboptimal under parameter uncertainty.

parameter arises, and if it does, it tends to be of smaller size. For δ=0.95 less than 1% of the simulations exhibit a bias after 20 periods and the average bias is reduced by more than half. A comparison of alternative values for the discount factor shows that the higher the discount factor, the more willing the agent will be to experiment, and thus reduce the likelihood and size of any bias which would arise under slow learning.

TABLE 5 ABOUT HERE

Table 5 reports the same measures as described above for payoff functions with different weights ω=(0.1,0.3,0.5) in the objective function. The percentage of simulations which exhibit a persistent bias under myopic behavior increases dramatically. For ω=0.5 a bias arises in one out of two simulations. Part of the reason for this increased likelihood of a bias due to mis-specified beliefs, is that positive values of ω are associated with an expanded range of incorrect limit belief and policy pairs, also including beliefs with *b* not equal to zero, as shown in section 4. The absolute size of the control and target biases in the second and third row of table 5 decreases as the weight ω increases, because higher values of ω imply a lower average value of the control variable and a higher average value of the target variable. This needs to be taken into account when interpreting the results for the average biases over all simulations in the fourth and fifth row of table 5. Although the percentage of biased simulations has increased, which tends to raise the average bias, the changes in the long-run averages of the control and target variable have an offsetting effect on the absolute value of the average control and target bias. Finally, a comparison of the third and seventh column shows that optimal learning results in much faster convergence to the truth and tends to reduce the control and target bias (where ω=0.3 and δ=0.75).

Clearly the results presented in tables 4 and 5 depend on the prior beliefs of the agent, the true underlying coefficient values and the variance of the shocks. I have conducted similar simulation exercises for alternative values of these parameters and have found that the generated data exhibits the same qualitative properties for a large range of reasonable parameter values. Reporting these results goes beyond the purpose of this paper. More importantly however, in practical applications of learning by doing it may be possible to obtain values for some of these input parameters directly from the data, while making the others the subject of sensitivity studies.

7. Conclusions

This paper provides a quantitative assessment of the value and extent of optimal experimentation in controlling a simple regression with two unknown parameters. Active experimentation is found to be optimal for a non-negligible range of prior beliefs and substantially improves expected future payoffs. I use an example with two possible parameter values to illustrate that there are discontinuities in the optimal policy for certain incorrect beliefs, because these beliefs would be self-reinforcing under myopic behavior. Perhaps not surprisingly, the extent of experimentation is largest in the neighborhood of these candidate incorrect limit beliefs. The same is true for a more general learning problem with continuous parameter space where the decision maker's beliefs are modeled as a bivariate normal distribution. This specification should be quite useful for economic applications, because from the agent's perspective the unknown parameters can take any real value.

Dynamic simulations reveal that myopic behavior neglecting the potential payoffs from experimentation frequently results in a bias in actions and outcomes. This bias arises because the decision maker holds on to mistaken beliefs about the unknown parameters and chooses actions that in turn reinforce these beliefs. Whenever the agent learns the truth after a period of mistaken beliefs, the change in behavior results in a shift in control and target variables. This generates nonstationarities in economic data that constitute observable implications of learning. Optimal experimentation instead leads to fairly rapid learning and steady convergence of observable outcomes towards equilibrium values along with a decline in variability.

The numerical approach in this paper can be used to study the link between learning by doing and experimentation in a variety of economic models. Interesting examples would be the impact of learning by doing on technology choice [see Jovanovich and Nyarko (1995) and (1996)] or the relative benefits of learning by doing and learning from others [see Foster and Rosenzweig (1995) with respect to agricultural production]. Another relevant example is optimal monetary policy under uncertainty about the relationship between policy instruments and targets. As shown in Wieland (1995) and (1996) learning has normative and positive implications for policy making after historical episodes of structural change.

Appendix

The numerical dynamic programming algorithm

The algorithm for computing the value function and optimal policy relies on successive application of the functional operator T which is based on the Bellman equation and defines a contraction mapping: (A.1)

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

$R(.)$ denotes expected current reward as a function of the current action x and the current state (belief) θ and is defined in table 1 and 2 respectively for each specification of beliefs. $w(.)$ refers to a continuous bounded function that is defined on the relevant state space Θ and constitutes an approximation of the value function V. δ denotes the discount factor and the integral represents the expected value in the next period, where next period's beliefs have been substituted out using the relevant updating equations $B(.)$ from tables 1 and 2. $p(\alpha,ß|\theta)$ is the normal density function implied by the agent's beliefs about the unknown parameters α and ß. Finally, $p(\epsilon)$ denotes the normal density function of the $N(0,1)$ error term.

We work with the space of continuous bounded functions mapping the state space Θ into the real line. This is a complete metric space in the sup metric (A.2)

$$d(w^0,w^1) = \sup_{\Theta} | w^0(\theta) - w^1(\theta) |$$

where w^0 and w^1 are continuous bounded functions. The operator T maps a continuous bounded function w into a continuous bounded function Tw. As shown by Kiefer and Nyarko (1989) Blackwell's sufficiency conditions, monotonicity and discounting, are satisfied and T is a contraction mapping such that (A.3)

$$d(Tw^1,Tw^0) \leq \delta\, d(w^1,w^0)$$

Thus T has a unique fixed point, the value function V, which can be calculated by value iteration, meaning successive application of T. $T^n w$ converges to V uniformly as $n \rightarrow \infty$. A convenient starting value w^0 is the single period reward function $R(.)$.

The algorithm is implemented as follows: first, calculate starting values w^0 for a grid of points in the state space Θ and save them in a table; second, calculate w^1 by applying the operator T to w^0 and update said table. This step involves a maximization with respect to the control x for each grid point in the state space. This maximization in turn requires repeated evaluation of the following integral: (A.4)

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

The updating equations B and the two normal density functions are known. The values of w^0

at the grid points can be read from the table and the values in between grid points can be approximated by multilinear interpolation.[16] The advantage of linear interpolation is that it preserves the shape of the function, positivity and monotonicity. Thus, even though the algorithm only remembers a discrete approximation of w_o, when computing values of w_1, linear interpolation guarantees that Blackwell's sufficiency conditions are satisfied and the algorithm remains a contraction mapping. Based on the updating equations, the density functions and the table w_o the above integral can be evaluated using Gaussian quadrature. The maximization step however, is nontrivial because there exist multiple local maxima. In fact, these multiple optima are what is behind the finding of non-differentiabilities in the value function and discontinuities in the optimal policy. To ensure that the global maximum is found, I first conduct a rough grid search, save the maximum, and then conduct a golden section search to compute the maximum more precisely. This two-step procedure is slow but secure. For each grid point in Θ this maximum is used to update the table approximating the value function, $w^1(.)$. The associated set of controls provides an approximation to the optimal policy, $x=h^1(.)$. This procedure is repeated to obtain w^2 and so on until the difference between two successive approximations is sufficiently small. A more detailed discussion of numerical dynamic programming, as well as the optimization and quadrature techniques utilized here can be found in Judd (1998).

Precision and computation costs

Since the algorithm is a contraction mapping, it is straightforward to construct a bound on the approximation error. If $w^{n+1} = Tw^n$, then $d(w^{n+1},w^n) \leq \delta\ d(w^n,w^{n-1})$ and after iterating i times $d(w^{n+1+i},w^{n+i}) \leq \delta^{1+i}\ d(w^n,w^{n-1})$. This implies the following upper bound on the approximation error: (A.5)

$$d(V,w^n) \leq \sum d(w^{n+1+i},w^{n+i}) \leq \frac{\delta}{1-\delta} d(w^n,w^{n-1})$$

This bound only depends on the discount factor and the maximal distance between the approximations obtained in the last two iterations. After every iteration, the algorithm checks all the grid points for the maximum of $|w^n(\theta)-w^{n-1}(\theta)|$ and computes the bound on the approximation error. The value iterations are stopped once the discounted relative difference between two successive approximations is within 0.5%. This error bound neglects that the numerical maximization and integration procedures as well as the linear interpolation of w between grid points are possible sources of approximation error. For this reason I have conducted a detailed sensitivity study for the bivariate normal specification with continuous parameter space. First, the final accuracy of the optimization routine is 0.01 percent rendering this step a negligible source of approximation error. Second, I have computed approximations using 10-, 15-, 20-, 30- and 100-point Gaussian quadrature, and find that increasing the number of points beyond 30 has negligible effects. Third, I have obtained approximations using successively greater numbers of grid points in each dimension. The density of the grid is most important in areas of high curvature such as

[16] Since we can use R or another known function as starting point w^0, the values in between grid points could be calculated exactly in the first iteration. However, in subsequent iterations, the functional form of w^n is not known and values in between grid points are obtained by interpolation.

along the slope *b*. The grid in this dimension is represented by the dots in figures 3 and 4. Comparing two value function approximations with 27 and 53 grid points for *b*, I find that the average percentage difference at the grid points is 0.33 percent, while the maximum percentage difference is 5.5 percent. The maximum difference in percentage terms is very small in absolute terms, since it occurs when the absolute value of the value function is very small. The average absolute difference is 0.014 and the maximum absolute difference is 0.4. These differences are small enough to be undetectable in the figures.

A drawback of the algorithm is that computational effort increases geometrically with the number of state variables. For example, if one uses N grid points for each dimension, the integration and optimization procedures described above have to be carried for each grid point i.e. N^5 times to complete one value iteration. The search for the optimum is especially time consuming, because of the existence of multiple local optima. An additional factor in terms of computational effort is that the number of value iterations required to achieve convergence within a set maximal error bound increases rapidly with the discount factor.

I have taken two different steps to reduce computation time. The first step was to introduce policy iterations, which reduces the number of value iterations required for convergence within a set maximal error bound. A policy iteration implies the application of the following operator: (A.6)

$$T^P w = R(h(\theta),\theta) + \delta \int w(B(\theta,x,\alpha+\beta\, h(\theta)+\epsilon))p(\alpha,\beta \mid \theta)p(\epsilon)\,d\alpha\,d\beta\,d\epsilon$$

Here $h(\theta)$ refers to the approximation of the policy function obtained from the preceding value iteration. Thus, following every value iteration, the operator T^P is repeatedly applied to the approximation of the value function *w* (for a given $h(.)$) until it converges within a given bound. Such policy iterations can be carried out fairly quickly because they do not involve an optimization step. As a result, the number of value iterations as well as overall computing time is reduced. This is particularly useful, when considering high values of the discount factor.

Secondly, for simple payoff functions such as the quadratic function with weight ω equal to zero, only four of the five state variables need to be considered and the optimization step has to be carried out only N^4 times per value iteration (the proof is given below). The approximations shown in figures 3 and 4 are based on a four-dimensional grid with 18, 53, 32 and 63 grid points respectively and use 30-point Gaussian quadrature. The grid points were chosen so that the distance between them is smaller in areas of high curvature. For this case, convergence as defined by a 0.5% maximal error was achieved after almost 2 days on a SPARC 20 work station. However, if the weight in the payoff function is positive, such as in column 4 of table 5 (δ=0.75,ω=0.3) a five-dimensional grid is needed. By using a slightly less dense grid and 15-point quadrature, a good approximation can be obtained within a week on the same computer.

A transformation that reduces the number of relevant state variables

One can ignore one state variable when dealing with the simple payoff function (A.7)

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

In this case the value function V corresponds to the supremum of

(A.8)

$$E\left[\sum_{j=0}^{\infty} -\delta^j (y_{t+j} - y^*)^2 \mid a_t, b_t, \Sigma_t\right] \quad \text{where} \quad y_j = \alpha + \beta x_j + \epsilon_j$$

Proposition: The value function $V(a,b,\Sigma)$ has the following property (for any k≠0): (A.9)

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

Proof: Consider the transformed problem

$$y_t = \alpha + \beta_k x_{k,t} + \epsilon_t \quad \text{where} \quad \beta_k = k\beta \quad x_{k,t} = x_t / k$$

If the prior on (α,ß) is $N(a_t, b_t, v_{a,t}, v_{b,t}, v_{ab,t})$, then the prior on β_k is $N(a_t, kb_t, v_{a,t}, k^2 v_{b,t}, kv_{ab,t})$. Since these two problems are equivalent, the proposition follows.
By setting $k = v_{b,t}^{-1/2}$ and replacing it in the above equation one obtains: (A.10)

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

Thus it is sufficient to approximate the value function for a grid of four state variables.

References

Aghion, P., P. Bolton, C. Harris, and B. Jullien (1991), Optimal learning by experimentation, Review of Economic Studies 58, 621-654.

Amman H., and D. Kendrick (1994), Active learning: Monte Carlo results, Journal of Economic Dynamics and Control 18, 119-124.

Amman H., and D. Kendrick (1995), Nonconvexities in stochastic control models, International Economic Review, 36 (2), 455-475.

Amman H., and D. Kendrick (1997), Erratum - active learning: A correction, Journal of Economic Dynamics and Control 21 (10), 1613-14.

Anderson, T., and J.B. Taylor (1976), Some experimental results on the statistical properties of least squares estimates in control problems, Econometrica 44, 1289-1302.

Balvers, R. and T. Cosimano (1994), Inflation variability and gradualist monetary policy, Review of Economic Studies 61, 721-738.

Balvers, R. and T. Cosimano (1993), Periodic learning about a hidden state variable, Journal of Economic Dynamics and Control 93, 805-827.

Balvers, R. and T. Cosimano (1990), Actively learning about demand and the dynamics of price adjustment, Economic Journal 100, 882-898.

Bar-Shalom, Y. (1981), Stochastic Dynamic Programming: Caution and Probing, IEEE Transactions on Automatic Control, AC-26 (5), 1184-1195.

Bertocci, G. and M. Spagat (1993), Learning, experimentation and monetary policy, Journal of Monetary Economics 32, 169-178.

Easley, D. and N. Kiefer (1988), Controlling a stochastic process with unknown parameters, Econometrica 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 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 103, 1176-1209.

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.

Jovanovich, B. and Y. Nyarko, (1996), Learning by doing and the choice of technology, Econometrica 64, 1299-1310.

Judd, Kenneth, 1998, Numerical Methods in Economics, (MIT Press, Cambridge, MA).

Kendrick, D. (1978), Non-convexities from probing an adaptive control problem, Economic Letters 1, 347-351.

Kendrick, D. (1981), Stochastic control for economic models, Economic Handbook Series, McGraw Hill, New York 1981.

Kendrick, D. (1982), Caution and probing in a macroeconomic model, Journal of Economic Dynamics and Control 13, 201-223.

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.

Lai and Wei (1982), Least Squares Estimates in Stochastic Regression Models with Applications to Identification and Control of Dynamic Systems, The Annals of Statistics 10, No. 1, 154-166.

McLennan, A. (1984), Price dispersion and incomplete learning in the long run, Journal of Economic Dynamics and Control 7, 331-347.

Mizrach, B. (1991), Non-convexities in a stochastic control problem with learning, Journal of Economic Dynamics and Control, 15, 515-538.

Norman A., (1976), First-order dual control, Annals of Economic and Social Measurement, 5 (3), 311-322.

Norman A., M. Norman and C. Palash (1979), Multiple relative maxima in optimal macroeconomic policy: an illustration, Southern Economic Journal 46, 274-279.

Prescott, E., (1972), The multi-period control problem under uncertainty, Econometrica 40, 1043-1058.

Nyarko, Y. (1991), The number of equations versus the number of unknowns: the convergence of Bayesian Posterior Processes, Journal of Economic Dynamics and Control, 15, 687-713.

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

Rustichini, A. and A. Wolinsky (1995), Learning about variable demand in the long run, Journal of Economic Dynamics and Control 19, 1283-1292.

Taylor, J.B., (1974), Asymptotic properties of multiperiod control rules in the linear regression Model, International Economic Review 15, 472-484.

Taylor, J.B., (1976), Methods of efficient parameter estimation in control problems, Annals of Economic and Social Measurement, 5, (3).

Trefler, D. (1993), The ignorant monopolist: optimal learning with endogenous information, International Economic Review 34 (3), 565-581.

Tse, E. and Y. Bar-Shalom, (1972), An actively adaptive control for linear systems with random parameters, IEEE Transactions on Automatic Control, AC-17, 38-52.

Tucci, M. (1996), The nonconvexities problem in adaptive control models: a simple computational solution, manuscript, University of Siena.

Tucci, M. (1997), Adaptive control in the presence of time-varying parameters, Journal of Economic Dynamics and Control 22 (1), 39-47.

Wieland, V. (1995), Optimal control with unknown parameters - a study of optimal learning strategies with an application to monetary policy, Ph.D. Thesis, Stanford University.

Wieland, V. (1996), Monetary policy, parameter uncertainty and optimal learning, manuscript, Federal Reserve Board, Washington, D.C.

Zellner, A., (1971), Introduction to bayesian inference in econometrics, New York, Wiley, 1971.

Table 1. Specification of the Illustrative Example

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

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

Myopic Policy

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

Bellman Equation

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

where f(.) is the predictive distribution of y

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

Table 2. Specification of the Learning Problem with Continuous Parameter Space

Updating Equations

$$\Sigma_{t+1} = \left[\Sigma_t^{-1} + X_t X_t \right]^{-1} \qquad \begin{pmatrix} a \\ b \end{pmatrix}_{t+1} = \Sigma_{t+1} \left[X_t' y_t + \Sigma_t^{-1} \begin{pmatrix} a \\ b \end{pmatrix}_t \right] \quad \text{where} \quad X_t = \begin{pmatrix} 1 \\ x_t \end{pmatrix}$$

Expected One-Period Reward

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

Myopic Policy

$$x^{my} = H^{my}(a,b,\Sigma) = -\frac{v_{ab} + (a - y^*)b}{v_b + b^2 + \omega}$$

Bellman Equation

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

Table 3. The Set of Possible Limit Beliefs and Limit Actions

Belief Invariance	$\bar{v}_a + \bar{v}_{ab}\bar{x} = 0 \quad \Rightarrow \quad a, v_a$ invariant $\bar{v}_{ab} + \bar{v}_b\bar{x} = 0 \quad \Rightarrow \quad b, v_b, v_{ab}$ invariant
One-Period Optimization	$\bar{x} = -\dfrac{(\bar{v}_{ab} + (\bar{a} - y^*)\bar{b})}{(\bar{v}_b + \bar{b}^2 + \omega)}$
Semi-Positive-Definiteness of Σ	$\bar{v}_a\bar{v}_b - \bar{v}_{ab}^2 \geq 0$
Non-Negativity of Variances	$\bar{v}_a, \bar{v}_b \geq 0$
Mean Prediction	$\alpha + \beta\bar{x} = \bar{a} + \bar{b}\bar{x}$

Table 4. Learning Biases under Myopic Behavior versus Optimal Learning

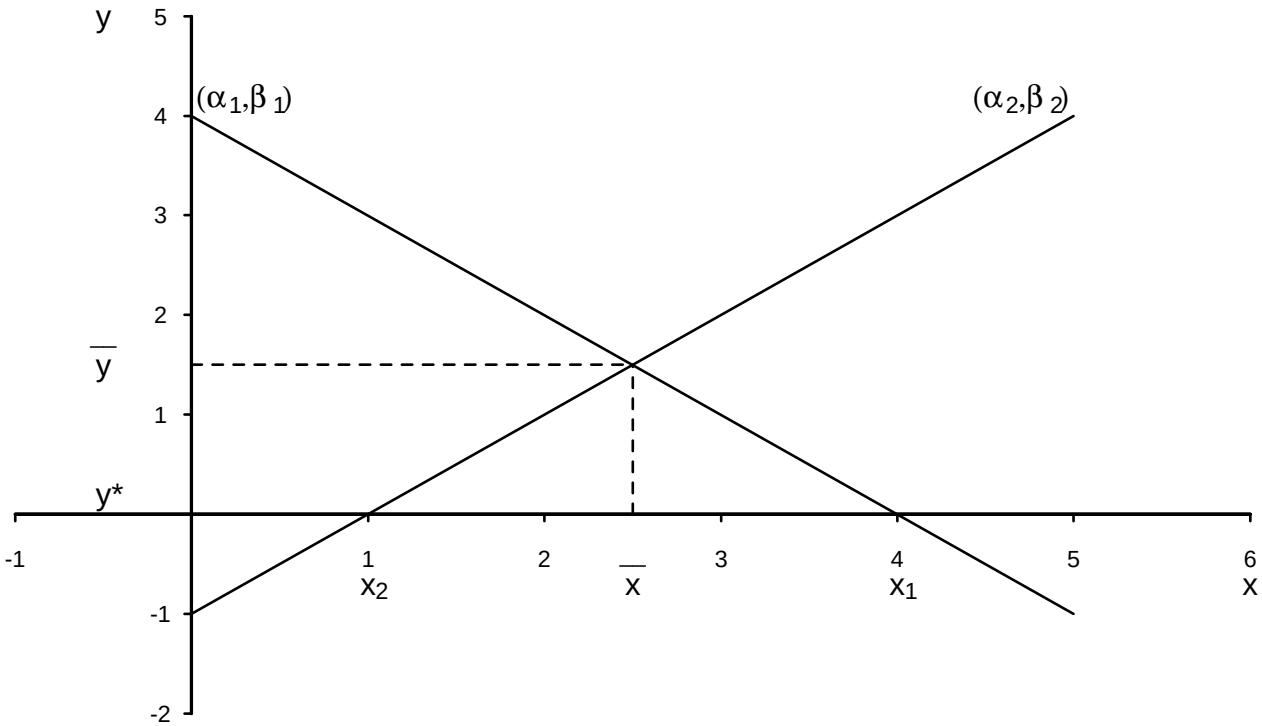
	Myopic Behavior		**Optimal Learning**					
			(δ=0.25)		**(δ=0.75)**		**(δ=0.95)**	
	T=20	*T*=40	*T*=20	*T*=40	*T*=20	*T*=40	*T*=20	*T*=40
% of Paths								
s.t. Bias>1 at *T*	11%	2.3%	6.9%	3.5%	1.1%	0.3%	0.8%	0.4%
Biased Paths:								
Control Bias	-2.55	-2.46	-2.16	-2.16	-1.39	-1.51	-0.83	-0.81
Target Bias	2.49	2.41	1.96	2.03	1.02	1.3	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

Note: These results are based on an objective function with the weight ω equal to zero.

Table 5. Learning Biases when Instrument Variation is Costly

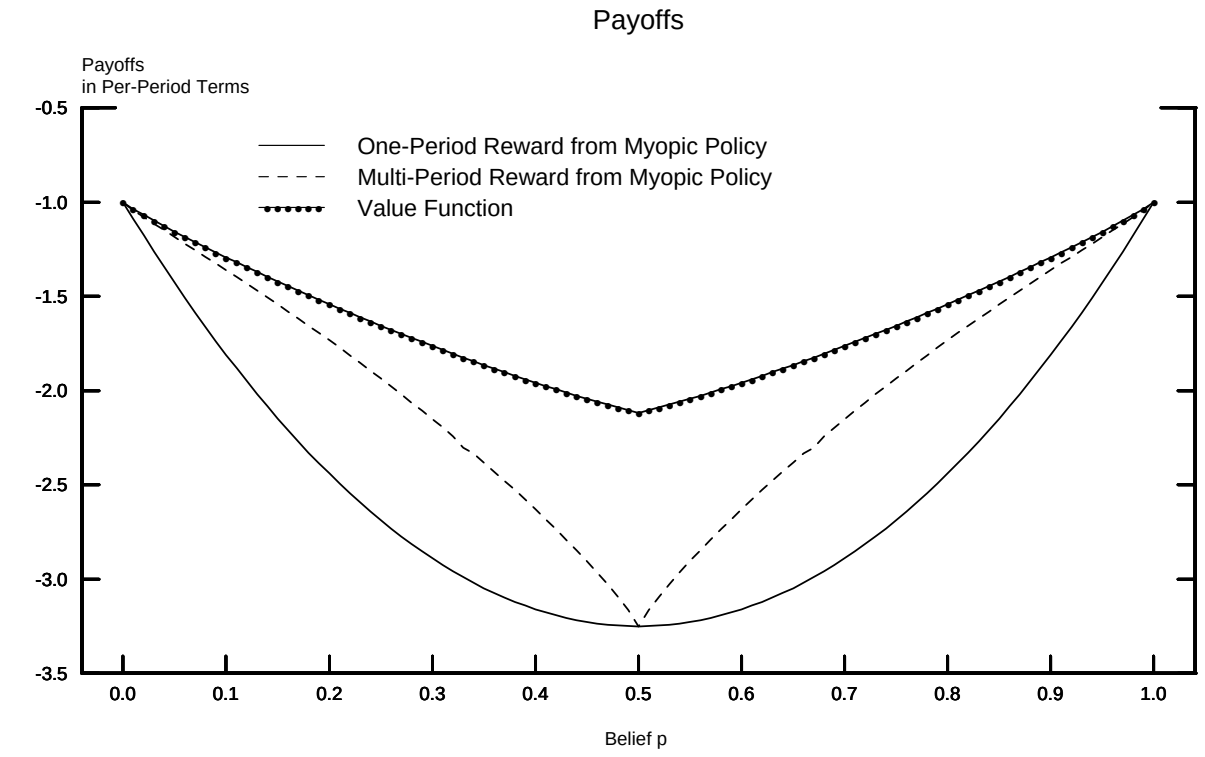
	Myopic Behavior						Optimal Learning	
	(ω=0.1)		**(ω=0.3)**		**(ω=0.5)**		**(ω=0.3,δ=0.75)**	
	T=20	*T*=40	*T*=20	*T*=40	*T*=20	*T*=40	*T*=20	*T*=40
% of Paths:								
Bias>1 at *T*	18%	8.4%	38%	25.4%	43.7%	33.1%	3.2%	2.7%
Biased Paths:								
Control Bias	-2.23	-2.07	-1.67	-1.58	-1.41	-1.38	-1.29	-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	-0.83	-0.57	-0.39
Target Bias	1.13	0.72	1.11	0.85	0.98	0.82	0.54	0.37

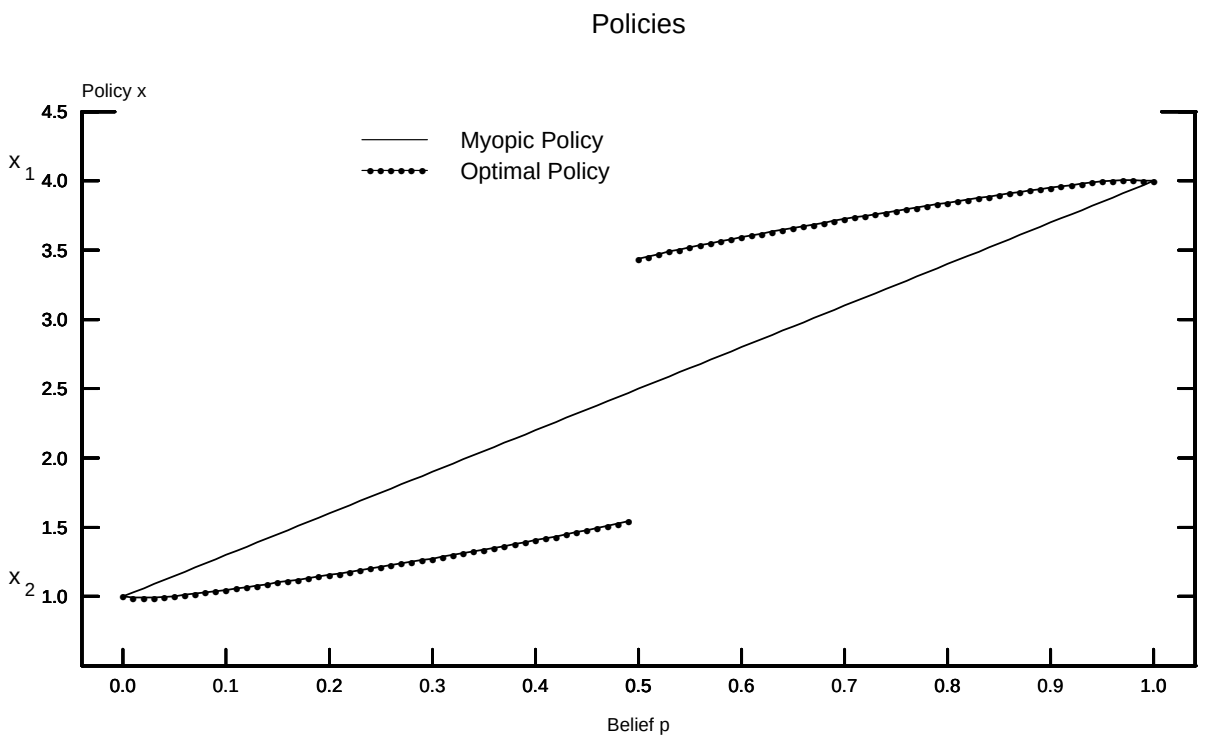
Figure 1 Illustrative Example with Discrete Parameter Space



Note: $\bar{x} = 2.5$ is an uninformative action which on average, would induce the same value $\bar{y} = 1.5$ under either of the two possible sets of parameter values, $(\alpha_1, \beta_1) = (4,1)$ or $(\alpha_2, \beta_2) = (-1,1)$.

Figure 2 Illustrative Example - Value and Policy Functions

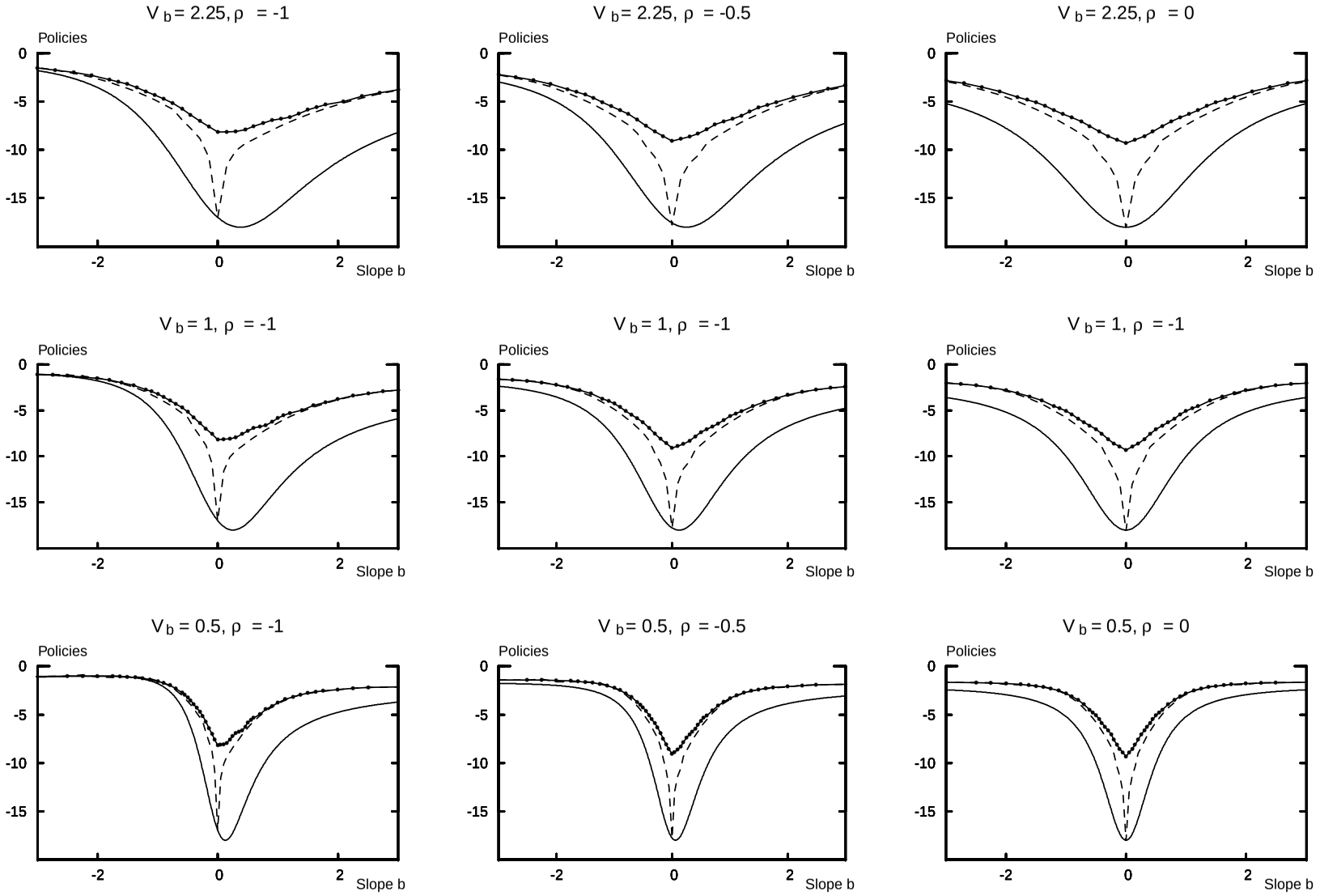




Note: The two possible sets of parameter values are $(\alpha_1,\beta_1) = (4,-1)$ and $(\alpha_2,\beta_2) = (-1,1)$. The discount factor δ is set equal to 0.75, the target value y* equals 0, and the preference parameter ω is set equal to 0.

Figure 3. Payoffs with Continuous Parameter Space

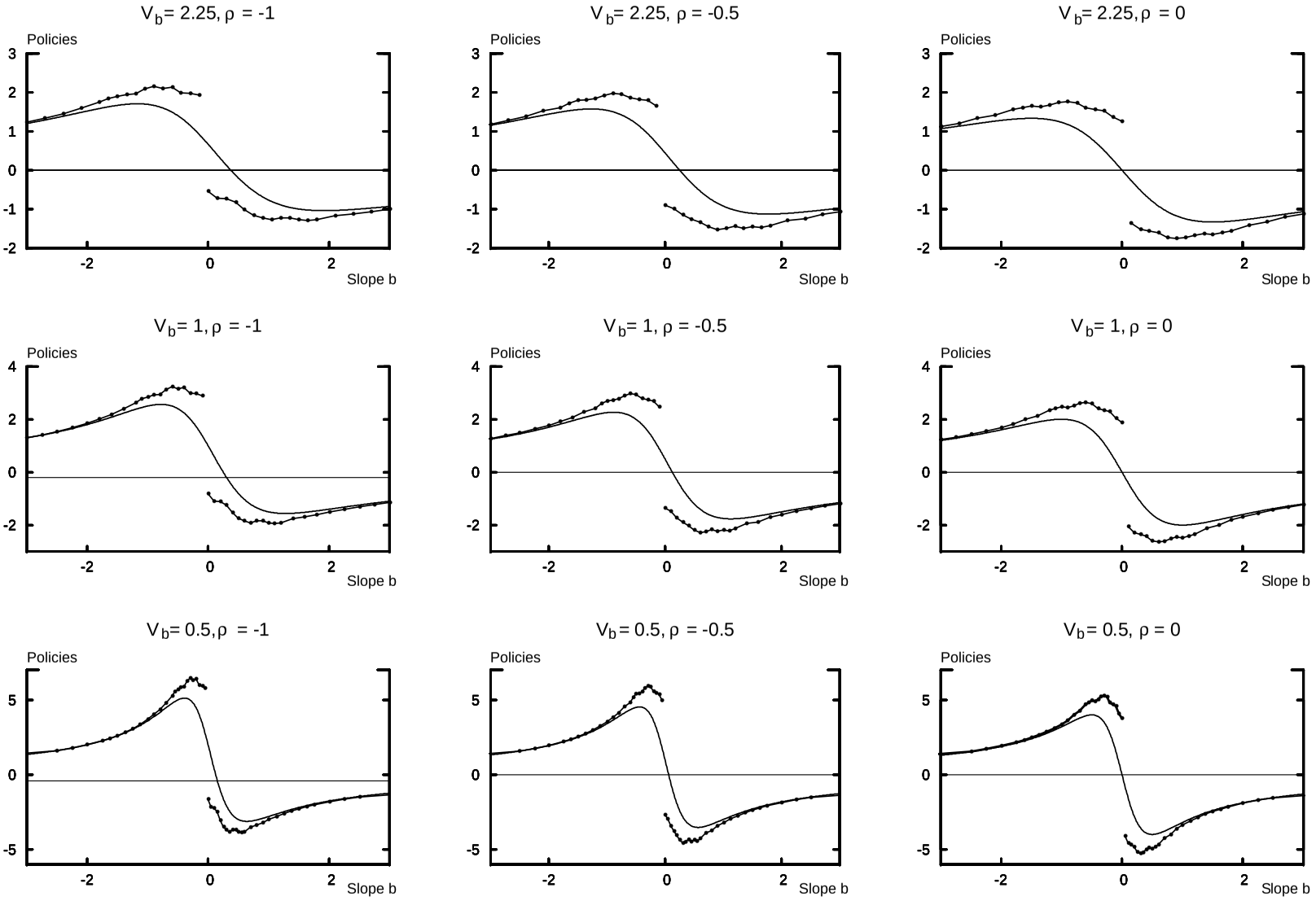
Value Function (dotted), One Period Reward (solid) and Multi-Period Reward (dashed) Under Myopic Policy



Note: In each panel v(a)=1 and a=4, the covariance is different in each panel and can be computed from the correlation coefficient ρ. The discount factor δ is set equal to 0.75, the target value y* equals 0, and the preference parameter ω is set equal to 0.

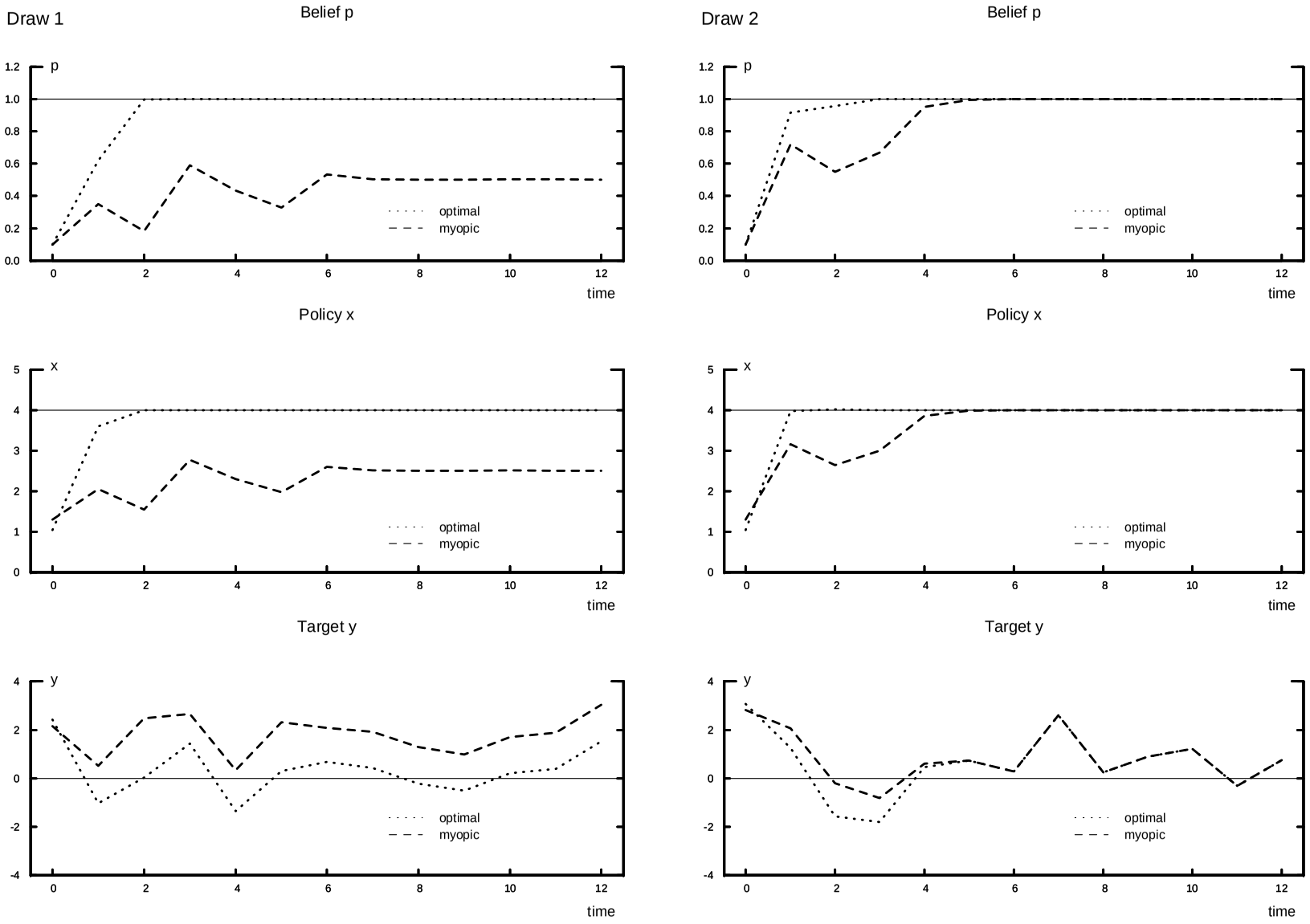
Figure 4. Policies with Continuous Parameter Space

Myopic Policy (solid) and Optimal Policy (dotted)



Note: In each panel v(a)=1 and a=4, the covariance is different in each panel and can be computed from the correlation coefficient ρ. The discount factor δ is set equal to 0.75, the target value y* equals 0, and the preference parameter ω is set equal to 0

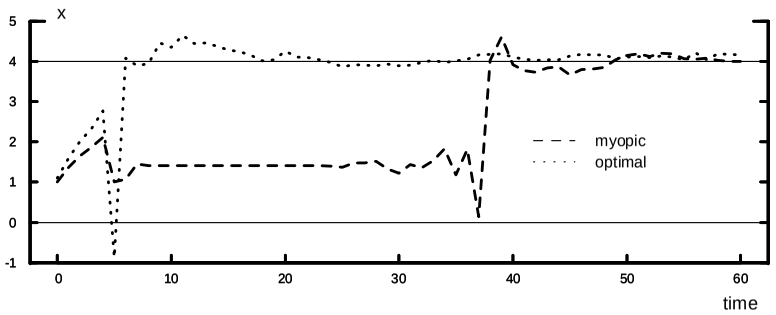
Figure 5 Illustrative Example - Dynamic Simulations



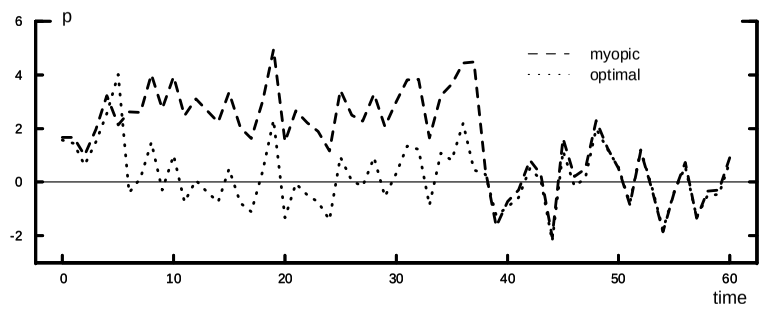
Note: The discount factor δ is set equal to 0.75, the target value y* equals 0, and the preference parameter ω is set equal to 0.

Figure 6 Dynamic Simulations with Continuous Parameter Space

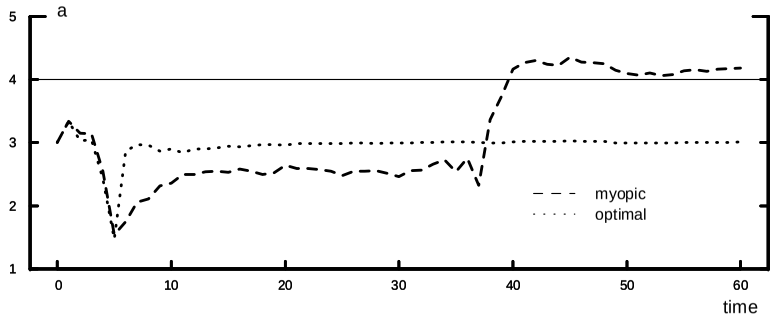
Control Variable x



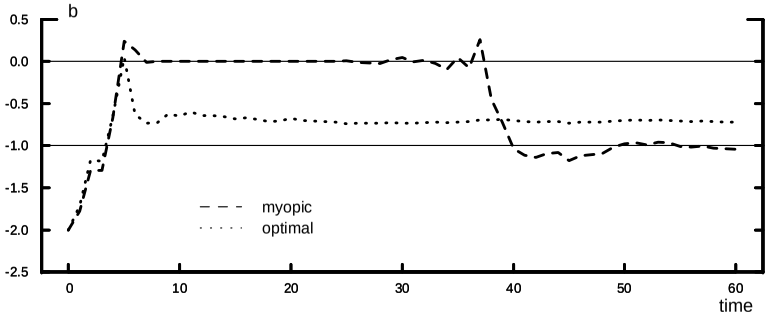
Target Variable y



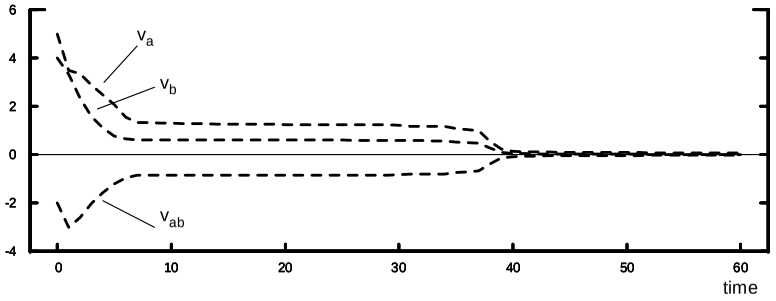
Beliefs: Mean of the Intercept a



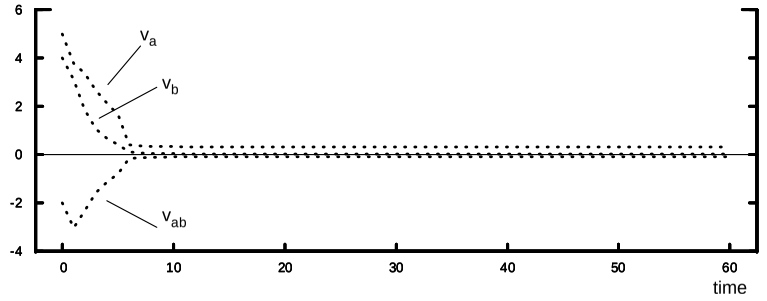
Beliefs: Mean of the Slope b



Beliefs: Parameter Uncertainty under Myopic Behavior



Beliefs: Parameter Uncertainty under Optimal Learning



Note: The discount factor δ is set equal to 0.75, the target value y* equals 0, and the preference parameter ω is set equal to 0.