Planning Costs and the Theory of Learning by Doing

W. Bentley MacLeod*
Department of Economics and The Law School
The University of Southern California
Los Angeles, CA 90089-0253
wmacleod@usc.edu

This version - August 2002
Keywords: decision theory, learning by doing, contingent plans

Abstract

Using Arrow (1958)’s distinction between a “choice” and a “decision”, this paper illustrates how the explicit introduction of planning costs into a model of decision making under uncertainty can result in a theory of learning by doing that is empirically implementable. The model provides a theoretical basis for an empirical learning curve first introduced by Newell and Rosenbloom (1981). The model is then fitted to data on corporate culture from Weber and Camerer (2001), and is shown to perform better for this application than a more standard Bayesian learning model.

* I would like to thank Boyan Jovanovic, Paul Rosenbloom, and Roberto Weber for supplying me with their data. The paper has benefited from the comments of George Akerlof, Colin Camerer, Janet Currie, Dick Day, Ido Erev, Phillip Jehiel, Yaw Nyarko, Bernard Sinclair-Desgagné and Joel Sobel. However, none are to be implicated for the deficiencies of the paper. Finally, I very much appreciated the research assistance provided by Mehdi Farsi. I am also grateful for the comments during presentations of earlier versions of this work at the American Economics Association Meeting, Chicago 1997, McArthur Foundation Conference, Hoover Institute, Stanford, August 1998, Conference on Cognition, Emotion and Rational Choice, Center for Comparative Political Economy, UCLA, April 2000, the Marshall School of Business, USC, and Munich University. The financial support of National Science Foundation grant SBR-9709333 is gratefully acknowledged, and I benefited from the hospitality of the Center for Economic Studies, Munich while working on this paper.
1 Introduction

The economic approach is built upon a model of human behavior that supposes individuals make choices that are consistent with a stable set of preferences. This approach, as exemplified by Becker (1976), has been very successful in explaining a wide range of economic phenomena, despite a growing body of evidence demonstrating that preferences can often be very unstable. For example, Tversky and Kahneman (1981) have shown individual that decisions can vary with factors that are irrelevant to the decision at hand, while Kahneman, Knetsch, and Thaler (1990) show that individuals exhibit endowment effects, namely preferences towards an object which appear to change as soon as they gain possession of the object.

This research, along with a large body of other work on anomalies in decision making, demonstrates that observed human behavior is extremely complex, and often inconsistent with predictions of utility theory.\footnote{See the collection Camerer, Loewenstein, and Rabin (2003). The introductory essay provides an nice overview of the current state of the art regarding anomalies in decision making.}

This leads one naturally to the question of why, despite these anomalies, the simple model of utility maximization has been so successful? Secondly, why do we observe individuals making simple errors in judgement in the laboratory, when at the same time these individuals are capable of very complex behaviors such as performing surgery, driving a car, or proving theorems, behaviors that are far beyond the capabilities of modern computers?

One approach to the problem, as suggested by Rabin (1998), is to modify the standard model of preferences in a direction that ensures greater consistency with the evidence. This approach faces two limitations. First, it does not directly explain why the simple utility maximization model has performed so well, and hence does not address the question of the appropriate scope of application for the standard model. Secondly, the introduction of a better descriptive model of preferences implies that the underlying forces of costs and benefits, fundamental for the economic approach, play less of a role in explaining observed behavior. Given that human behavior has evolved in response to challenges of surviving in a very harsh and competitive environment, one would expect that there is an underlying economic reason why it has been so successful.

The purpose of this paper is to explore a simple economic model of decision making with planning costs. The main result is that learning by doing, or more generally adaptive learning, is part of an optimal decision procedure. There is long tradition of using adaptive learning models in economics, though typically such behavior is viewed as the result of “bounded rationality”.\footnote{The adaptive learning model in economics was introduced by Simon (1955). Recent work work has focussed upon applications in game theory, and include Stahl (1996), Erev and Roth (1998), Camerer and Ho (1999) and Costa-Gomes, Crawford, and Broseta (2001). See Sobel (2000) for review of this literature.} In contrast, in this model learning by doing is derived as an optimal Bayesian decision rule in the presence of costly contingent planning. Moreover, the
convergence result is very robust, and does not depend upon the fine details of the problem, other than supposing that the environment satisfies some stability properties (a necessary condition for any model of learning).

The model builds upon Arrow (1958)'s distinction between a \textit{choice} and a \textit{decision}.\footnote{See Arrow (1958), page 1.} A simple example may help illustrate the difference. It is very common for neophyte teachers to talk towards the blackboard, and write in such a way that the students cannot see what is being written. If we were to analyze such behavior from the perspective of revealed preference theory one would have to conclude that the teacher consciously does not wish to communicate with the students. While in some cases this may be true, it is more likely that once it has been pointed out to the individual that this is occurring, then the individual can and does modify behavior to improve performance. In other words, though the instructor has made a \textit{choice} to talk to the blackboard, it is unlikely that this is the outcome of a deliberate decision that considers all the consequences of her actions (including low teaching evaluations). If the individual spends a great deal of time thinking about teaching (as is done in a business school), then she would make \textit{decisions} that are consistent with the goal of maximizing the effectiveness of her teaching.

The model introduced in the next section supposes that each period the individual faces the following sequence of decisions and choices. At the beginning of the period she spends time making a contingent plan in anticipation of an event that will require an immediate response. This period corresponds to “decision making”. A random event then occurs requiring a response that corresponds to a “choice” in the sense used above. That is, the individual is aware of her choice, but does not have sufficient time to explore all the implications of the choice, and hence it may not necessarily be optimal. Choices are linked to decisions by supposing that if an the individual has made a contingent plan for an event that subsequently occurs, or has previously experienced the event (and hence learned by doing), then the choice is optimal and consistent with her preferences.\footnote{The difference here is based upon Alan Newell (1990)’s distinction between decision making in the “rational band” - where contingent planning is possible - and choices in what he calls the “cognitive band” - the individual is aware of the choice made, but contingent planning is not possible in the time allotted.} In the absence of a plan or previous experience with an event, it is assumed there is a chance the individual chooses an action that is inconsistent with her true preferences - in other words the individual makes a mistake.

Given that planning for all possible contingencies is costly, a trade-off arises. An individual can either make a complete contingent plan, or save upon planning costs by using an incomplete plan and hope that she has included in her plan events that are likely to occur. If an unplanned for event occurs, then there is a chance she will make a mistake. However, should that event occur again in the future then she is assumed to
learn from her experience and make an optimal choice at that time. I follow Rothschild (1974), and model uncertainty by assuming individuals have beliefs over the possible probability distributions determining the likelihood of different events occurring. With experience individuals not only learn how to respond to particular events, they also learn which events are more likely. It is shown that as uncertainty increases then the amount of planning that is optimal decreases, and in the limit it may be optimal to engage in no planning.

This result is different from a standard Bayesian learning by doing model, such as Jovanovic and Nyarko (1995), in which an individual receives a noisy signal of a small number of unobserved parameters each period. More precise information regarding these parameters results in increased performance. Hence this class of models also predicts a learning curve in which performance increases with experience. However these models do not explicitly address the trade-off between a decision and a choice. In section 3 it is shown that this is not merely a semantic distinction, but that in fact the two models have different testable empirical implications.

The agenda of the paper is as follows. The next section introduces the basic model, and provides conditions under which “learning by doing” is an optimal procedure. Section 3 discusses the empirical implementation of the procedure, and shows that the model provides a theoretical foundation for a class of learning curves, first suggested by Newell and Rosenbloom (1981), that are appropriate for modelling a wide variety of learning data. The empirical potential of the model is illustrated using data on corporate culture taken from Weber and Camerer (2001), and it is shown to provide a better fit than the more standard Bayesian learning model studied by Jovanovic and Nyarko (1995). The purpose of section 4 is to discuss the implications of the model for behavioral economics, and how the model can be used to reconcile utility maximization with the existence of anomalies in decision making. Section 5 contains a brief concluding discussion.

2 The Model

Consider an individual making choices at times $t = 0, 1, 2...$ given by $d_t \in D$ in response to an event $\omega^t \in \Omega$, where $\Omega$ is a finite set of events. Without loss of generality the decision set is restricted to a binary choice: $D = \{0, 1\}$. In addition, it is assumed that the events are selected by an i.i.d. stochastic process, where $\mu(\omega)$ is the probability that $\omega$ is chosen in period $t$. The decision maker does not know this distribution, but updates

---

5 If the set $D$ consists of $n$ decisions, then simply identify each $d$ to a binary number and hence one can set $D = \{0, 1\}^k$, where $k$ is the smallest integer such that $k \geq \log_2 n$. One then applies the analysis to a vector of decision functions, where each component takes values in $\{0, 1\}$. 

---
beliefs over the set of possible prior distributions as events are observed. Suppose that the agent’s utility is $U(d|\omega)$ if decision $d$ is chosen when event $\omega$ occurs. Throughout, it is assumed that $U(1|\omega) \neq U(0|\omega)$, hence for each event $\omega$ there is a unique optimal choice given by $\sigma^*_\omega = \arg\max_{d\in D} U(d|\omega)$.

It is assumed that the decision is sufficiently complex that there is not sufficient time between the observation of $\omega$ and the choice $d$ for individuals to consistently choose the optimal strategy. Before facing the event the individual may prepare a response either through an explicit contingent plan or by training with different possible events before facing a decision. For example, airline pilots train with flight simulators to prepare responses for various possible aircraft failures. Though the pilot may be capable of deducing the appropriate response for a particular failure, he or she may not have sufficient time to carry out such an analysis in the face of an actual failure during flight. Indeed, the point of pilot training is to ensure that the appropriate decision is taken quickly with little apparent thought.

The amount of planning is endogenous, and is the set of events for which the individual has a prepared response in period $t$, denoted by $\Omega^t \subset \Omega$. This implies that if $\omega^t \in \Omega^t$ occurs then the individual is able to respond optimally with $\sigma^*_\omega(\omega^t) \in \{0, 1\}$. If $\omega^t \notin \Omega^t$, then the individual does not have a prepared response, nor does she have sufficient time to determine the appropriate response, and hence she randomizes over $D$. For simplicity suppose that when the optimal response is made the reward is $u_g$, while if a plan is not in place for an event the expected payoff is $u_b < u_g$.\footnote{That is $u_b = (U(0|\omega) + U(1|\omega))/2$.} The choice of an individual at time $t$ is given by the function:

$$\sigma_t(\omega^t) = \begin{cases} \sigma^*_\tau(\omega^t), & \text{if } \omega^t \in \Omega^t, \\ \{\frac{1}{2}, \frac{1}{2}\}, & \text{if } \omega^t \notin \Omega^t. \end{cases}$$

where $\{\frac{1}{2}, \frac{1}{2}\}$ denotes the lottery that selects each action in $\{0, 1\}$ with equal probability. The function $\sigma^*_\omega(\omega^t)$ is the optimal response to $\omega^t$ at date $\tau$, where $\tau$ is the most recent time at which the agent either constructed a plan for $\omega^t$, or had experienced $\omega^t$. The requirement of using the most recent time $\tau$ is not necessary for the current section, but plays an important role in the subsequent sections with non-stationary dynamics.

Events are added to the set $\Omega^t$ in two ways. First there is learning by doing, if $\omega^t$ occurs then the agent evaluates her performance ex post, and encodes an optimal response to the event $\omega^t$, which is then added to the set $\Omega^t$. The second method is through the explicit formation of a contingent plan. The individual can expend effort before the realization of $\omega^t$ to add additional events to the set $\Omega^t$. This is assumed to be a costly activity, either because acquiring the behavior requires expensive training, or simply because of the cost associated with adding a large number of contingent plans. The goal then is to explicitly model the
trade-off between learning by doing and the formation of a contingent plan *ex ante*.

It is assumed that the individual knows that the events in \( \Omega \) are generated by some unknown stationary distribution \( \mu \in \Delta^N \). Since there are a finite number of events, it is possible to allow completely non-parametric beliefs, namely any distribution in \( \Delta^N \). A convenient prior belief for this situation is given by the Dirichlet distribution, \( f (x | \alpha) \), where \( \alpha = \{\alpha_1, ..., \alpha_N\} \). This distribution forms a conjugate family, which means that each period the updated beliefs after observing a draw from this multinomial distribution also has a Dirichlet distribution (see DeGroot (1972)). The parameter \( \alpha_i \) represents the weight associated with the event \( \omega_i \).

If event \( \omega_i \) is observed in period \( t \), then the posterior belief of the individual using Bayes rule is given by a Dirichlet distribution with parameters

\[
\alpha_t^j = \alpha_t^{j-1} \quad \text{if} \quad j \neq i \quad \text{and} \quad \alpha_t^i = \alpha_t^{i-1} + 1,
\]

where \( \alpha_t^{i-1} \) is the period \( t \) belief parameter. The expected value of the probability that event \( \omega_i \) occurs given a Dirichlet distribution with parameter \( \alpha \) is

\[
\alpha_i \sum \alpha_j.
\]

It is assumed that initial beliefs are given by \( \alpha^0 = b (\lambda_1, ..., \lambda_n) \), where \( \sum \lambda_i = 1 \). The parameter \( b \) defines the degree of uncertainty regarding beliefs, and determines the speed at which beliefs are undated. When \( b \) is very large, then the individual is very certain that the true distribution is close to \( \{\lambda_1, ..., \lambda_n\} \in \Delta^N \), and will update beliefs away from this distribution very slowly. In contrast when the agent is very uncertain and \( b \) is close to zero, then after one period the individual believes that the probability of an event from \( \Omega^t \) occurring is close to one. The parameter \( \lambda_i \) defines how the agent’s initial beliefs vary over the set of possible events. One special case that will be considered is when the decision maker has prior beliefs given by the principle of insufficient reason, namely \( \lambda_i = 1/N \), and initial beliefs are given by \( \{b/N, ..., b/N\} \) (see Jeffreys (1948)).

The cost of a contingent plan is assumed to be an increasing function of the number of states, and hence planning may be spread over several periods. At the beginning of period \( t \), but before observing \( \omega^t \), the individual may decide to make contingent plans for an additional \( n^t \) events in \( \Omega \setminus \{\Omega^{t-1} \cup \{\omega^t\}\} \), (it is assumed that the event that occurs in period \( t - 1 \) is always added to \( \Omega^t \)). The cost of this additional contingent planning is \( c(n^t) \), where \( c(0) = 0, c' > 0, c'' \geq 0 \). Let the marginal cost of going from \( n^t - 1 \) to \( n^t \) be given by

\[
m\cdot c(n^t) = c(n^t) - c(n^t - 1).
\]

The individual chooses the amount of planning in period \( t \) to maximize her discounted expected payoff given her beliefs. The benefit from adding a plan for event \( \omega \) arises from a net gain of \( u_g - u_b \) that accrues the first time \( \omega^t \) is observed times the probability that this event occurs. This gain is traded off against the cost of including that state in a contingent plan.

---

\[\text{If } \alpha_i \text{ goes to } \infty \text{ while the other parameters remain fixed then this distribution approaches the measure that places probability } 1 \text{ on the } i'^{\text{th}} \text{ event.}\]
Events that are not in $\Omega^t$ have never occurred in the past, and hence the expected probability that an event $\omega_i$ not in $\Omega^t$ occurs in period $T > t$, given that it has not occurred before, is:

$$\pi(t, T, b, \lambda_i) = \prod_{n=0}^{T-1} \left( 1 - \frac{b\lambda_i}{b+t} \right)^{\delta^n} \frac{b\lambda_i}{b+T}.$$ 

where as a matter of convention $\pi(t, t, b, \lambda_i) = \frac{b\lambda_i}{b+t}$. Due to the Bayesian updating, for any event that has not been observed, the probability that such an event will occur decreases with time. Since this probability does not depend upon the events that have occurred, and since we have assumed that the distribution is an i.i.d. process then the benefit from adding a plan for an event $\omega$ that has not been observed is history independent. The marginal benefit from adding an event $\omega_i$ to the set $\Omega^t$ that has not been observed in the past is given by:

$$mb(t, b, \lambda_i, \delta) = (u_g - u_b) \sum_{n=0}^{\infty} \pi(t, t + n, b, \lambda_i) \delta^n. \quad (2)$$

The first proposition characterizes the properties of the marginal benefit function.

**Proposition 1** For $t \geq 0$, $b > 0$, $\lambda_i, \delta \in (0, 1)$, the marginal benefit function $mb(t, b, \lambda_i, \delta)$ is strictly decreasing in $t$ and strictly increasing in $b$, $\lambda_i$ and $\delta$. In addition the marginal benefit function has the following limit values:

$$\lim_{b \to \infty} mb(t, b, \lambda_i, \delta) = (u_g - u_b) \frac{\lambda_i}{1 - \delta + \lambda_i \delta}$$

$$\lim_{b \to 0} mb(t, b, \lambda_i, \delta) = \begin{cases} (u_g - u_b) \lambda_i & \text{if } t = 0, \\ 0 & \text{if } t > 0. \end{cases}$$

The proof of this proposition is in the appendix. The intuition is for the result is as follows. The marginal benefit from planning for a particular event $\omega_i$ is given by the marginal value of a good decision $(u_g - u_b)$ times the discounted probability of this event occurring. With time, if the event has not occurred, then one believes it is less likely occur and therefore the benefit from planning falls. The parameter $\lambda_i$ is the ex ante probability of this event occurring and hence the benefit from planning increases if the event is thought to be more likely. In the case of $b$, an increase in $b$ implies that one believes are less sensitive to new information, and hence the fact that event $\omega_i$ has not occurred has a smaller downward impact on beliefs, and hence the benefit from planning is greater. Increasing $\delta$ implies that the value from the event occurring in the future is higher, and hence the gains from having a plan in place are greater.

The first limiting case corresponds to the case that the individual knows for sure that the probability of event $\omega_i$ occurring is $\lambda_i$. In this case one can immediately see that marginal benefit from planning is always
positive, and constant with time, with an upper bound of \((u_g - u_b)\) that is achieved when \(\delta = 1\). The case when \(b\) is close to zero corresponds to almost complete uncertainty regarding the true distribution. As a consequence after the first period beliefs assign close to probability one to the set \(\Omega^t\), the events that have already occurred. Hence, it immediately follows in this case that for any events not included in a contingent plan in the first period, they will never be included in any future plan.

The fact that the marginal benefit falls with \(t\) implies that if one should implement a plan at date \(t\), then the benefit of implementing it earlier is higher. When planning costs are linear this implies that the optimal decision has the following simple structure.

**Proposition 2** Suppose that the marginal costs of planning are constant, \(mc(n) = c\) for all \(n\), then given \(b\) and \(\delta\), there is a belief \(\lambda^* (b, \delta)\) such that individuals form a contingent plan at date \(t = 0\) for all events \(\omega_i\) such that \(\lambda_i \geq \lambda^* (b, \delta)\). For all remaining events and dates all learning occurs via experience rather than planning.

**Proof.** Since \(mb(t, b, \lambda_i, \delta) \geq mb(t + 1, b, \lambda_i, \delta)\), then \(mb(t, b, \lambda_i, \delta) - c > \delta (mb(t + 1, b, \lambda_i, \delta) - c)\), and it is never optimal to delay a planning. Define \(\lambda^* (b, \delta)\) by \(mb(0, b, \lambda^* (b, \delta), \delta) = c\), and since \(mb(t, b, \lambda, \delta)\) is strictly increasing in \(\lambda\) it follows that for all \(\lambda_i \geq \lambda^* (b, \delta)\) it is optimal to make a contingent plan for all events such that \(\lambda_i \geq \lambda^* (b, \delta)\).

Thus the individual makes contingent plans for the most likely events, as long as the marginal costs of planning are not too high. In the extreme case when \(c\) is sufficiently high the individual’s optimal strategy is to make no contingent plans, and improve performance only through learning by doing. Conversely, with sufficiently low marginal costs, complete contingent planning is optimal. Notice that for any positive \(c\), one can bound the fraction of states for which an individual will make a contingent plan.

**Corollary 3** The maximum number of states for which an individual would construct a plan is bounded by \(\bar{N} = \text{int} \left( \frac{1}{\lambda^* (b, \delta)} \right)\), and hence for any \(N > \bar{N}\) planning is necessarily incomplete.

Therefore, when the number of states is sufficiently large, all other things being equal, planning is necessarily incomplete.

Consider now the implications of assuming the marginal cost of planning is increasing with the number of states. In this case individuals may spread planning over several periods. To simplify the analysis the principle of insufficient reason is used to construct the prior distribution, and hence \(\lambda_i = 1/N\), and \(N\) is now a parameter of the model. The extension to a non-uniform prior is straightforward, and the optimal procedure with have the same generic structure, with more likely events included in any plans before less likely events.
Proposition 4 Suppose that cost of planning is strictly convex and increasing in \( n \), then if the number of events \( N \) is sufficiently large then the optimal level of planning, \( n^\ast(t,b,N) \), satisfies:

\[
mc(n^\ast + 1) \ge mb(t, b, 1/N) \ge mc(n^\ast).
\]

Moreover, the optimal amount of planning is decreasing with time, and is increasing with one’s certainty regarding the true distribution (greater \( b \)). If the number of states is greater than \((u_g - u_b)/mc(1)\) then when the individual’s beliefs are sufficiently uncertain (\( b \) sufficiently small) the individual makes no contingent plans, and only learning by doing is optimal.

This result illustrates the basic trade-off between planning and learning by doing. In particular when the event space is sufficiently complex, and beliefs are sufficiently uncertain, then it is not optimal to make any contingent plans. The next proposition considers the optimal strategy in the case that marginal costs of planning are rising, and the state space is sufficiently small that the individual would in finite time have in place a complete contingent plan. However, due to the rising marginal costs of planning in a particular period the individual may spread planning over several periods.

Proposition 5 Suppose planning costs are increasing and convex in \( n \), then there exists an optimal planning rule \( n^0(r_t,t,b,N) \), where \( r_t \) is the number of unexplored states given by \( \Omega \setminus \Omega^{t-1} \). This rule has the following properties:

1. \( n^0(r_t,t,b,N) \) is increasing in \( r_t \).
2. If \( r_t \leq \tilde{n}(t,b,N) \), then \( n^0(r_t,t) = r_t \).
3. If \( r_t > \tilde{n}(t,b,N) \), then \( \min\{n^*(t,b,N), r_t\} \geq n^0(r_t,t,b,N) \geq \tilde{n}(t,b,N) \), where \( n^*(t,b,N) = \arg\max_{n \geq 0} n \cdot mb(t,b,N) - c(n) \) is the optimal one period strategy and \( \tilde{n}(t,b,N) \) satisfies:

\[
mb_t - mc(\tilde{n}) \geq \max\{p_{t+1}\delta(mb_{t+1} - c(1)), 0\} \geq mb_t - mc(\tilde{n} + 1),
\]

where \( p_t = (N + t/b)^{-1} \) is the probability that an event that has not been observed occurs in period \( t \), and \( mb_t = mb(t,b,1/N) \).

Hence, an optimal strategy exists regardless of the number of events, and that the amount of planning is increasing with the number of unexplored states. In other words the pattern of planning is the same as in the previous case, except that now the individual may choose to explore less than \( n^*(t,b,N) \) events due to the option value of waiting a period to finish making her contingent plan.
2.1 Uncertain Choice

It has been assumed that conditional upon an event there is a unique optimal choice that does not vary over time. Consider now the situation in which there is a chance that the optimal strategy varies each period. The introduction of uncertainty regarding choice is important for many decisions, for example, when deciding to whether or not to sell an asset one’s decision is sensitive to whether or not one expects the price to rise or fall in the next period. The model can be extended to deal with such cases by letting each

\[ \omega = \{ \omega_0, \omega_1 \} \]

with the interpretation that if \( \omega_0 \) occurs, then choosing action 0 is optimal, while if \( \omega_1 \) occurs choosing 1 is optimal.

It is assumed that in period \( t \) the individual observes the event \( \omega_t = \{ \omega^t_0, \omega^t_1 \} \in \Omega \) before making a decision, but not \( \omega^t_0 \) or \( \omega^t_1 \). After the decision is made the individual learns which would have been the optimal choice. Further, suppose that the probabilities evolve according to the following rule:

\[
\Pr \{ \omega = \omega^t_0 | \omega^0, \omega^1, ..., \omega^{t-1} \} = \begin{cases} 
a_{\omega_0} & \text{if } \omega_0 \text{ occurred last time event } \omega^t \text{ occurred,} 
(1 - a_{\omega_1}) & \text{if } \omega_1 \text{ occurred last time event } \omega^t \text{ occurred,}
\end{cases}
\]

(4)

Notice that if \( a_{\omega_0} = a_{\omega_1} = 1 \) then this is exactly the process studied above in which the optimal strategy does not change over time. When \( a_{\omega_0}, a_{\omega_1} > 1/2 \) then the process exhibits persistence, that is if \( \omega_0 \) occurred previously, then it is more likely to occur again when \( \omega^t \) occurs.

Finally, one needs to consider the gains from planning in the context of this model. Suppose that the parameters \( a_{\omega_0}, a_{\omega_1} \) are not known, but in period \( t \) if the agent plans for event \( \omega^t \), then she will know the optimal choice for that period. Due to the introduction of uncertainty, the gains from planning are reduced, and one has the following proposition.

**Proposition 6** Suppose the events follow the Markov chain given by 4 and the agent believes \( a_{\omega_0}, a_{\omega_1} \in (1/2, 1) \). Then if \( c > mb(0, b, 1/N) \) (as defined by 2) learning by doing is an optimal Bayesian decision procedure.

**Proof.** If event \( \omega_0 \) occurs, when \( \omega \) is observed, then the next time \( \omega \) occurs it is optimal to choose \( d = 0 \) because the probability of 0 being optimal is \( a_{\omega_0} > 1/2 \). However, the marginal benefit from planning is less than \( mb(0, b, N) \) because \( a_{\omega_0} \leq 1 \), and hence the result follows immediately from proposition 2.

This proposition captures the basic normative feature of the learning by doing algorithm. In situations for which the optimal choice in the past for a given event remains optimal in the future, then learning by doing is optimal in a complex environment. Moreover, behavior is adaptive, that is if in a given period the optimal choice changes, then the individual’s behavior also changes to this choice in the future. It highlights
one of the desirable features of learning by doing, namely that behavior can evolve to conform to the optimal choice in a non-stationary environment, as long as the environment does not change too quickly, as captured by the assumption \( a_i > 1/2 \).

3 Learning Dynamics

The purpose of this section is to derive the learning curve implied by this model, and to compare it to a standard one parameter learning model due to Jovanovic and Nyarko (1995) using an interesting data set on learning from Weber and Camerer (2001). Only the simplest possible variant of the model is considered, and it is assumed that the parameters \( b \) and \( c(\cdot) \) are such that planning is never optimal, and hence the individual engages in only learning by doing.\(^8\) Suppose that there are \( N \) possible events and the true probability of \( N \) occurring is \( 1/N \). This is consistent with some planning occurring as long as all planning occurs in the first period, and for all the remaining events the individual uses the principle of insufficient reason to assign unobserved events equal probabilities. At time \( t \) let \( m_t \in \{0, 1, \ldots, N\} \) be the number of events that have been experienced, which, given the equal probability hypothesis, is a sufficient statistic for the state of an individual’s knowledge. In period \( t \) if an event from these \( m_t \) events occurs then the individual earns \( u_g \), otherwise the payoff is \( \hat{u} = (u_g + u_b)/2 \). Hence the expected payoff in period \( t \) is:

\[
V(m_t) = u_g \cdot m_t/N + u_b \cdot (1 - m_t/N) .
\]

Whenever a new event is experienced this is added to the repertoire \( \Omega_t \), and \( m_t \) is increased by 1. The evolution of the state over time is a Markov chain with transition probability function:

\[
P [m_{t+1}|m_t] = \begin{cases} 
  m_t/N, & \text{if } m_t = m_{t+1}, \\
  1 - i/N, & \text{if } m_t = m_{t+1} + 1, \\
  0, & \text{in all other cases.}
\end{cases}
\]

In the data one observes neither \( N \) nor the number of events experienced. Moreover, the data reports the average performance for a number of individuals, rather than actual performance after \( t \) trials. Thus the expected payoff is used as the performance measure in the model, \( x^t_m \) is the probability that an individual has experienced \( m \) events in period \( t \), and \( x^t = (x^t_0, x^t_1, \ldots, x^t_N) \) is the vector of probabilities. The initial state is assumed to be \( x_0 = (1, 0, \ldots, 0) \), which assigns probability 1 to having no experience. Let \( P \) denote the Markov transition matrix, where \( P_{ij} \) is the probability of going from state \( i \) to state \( j \), then the probability

\(^8\)Alternatively, one may suppose the current model is the reduced form after all planning has occurred.
distribution of experience at date \( t \) is given by:
\[
x^t = x^0 P^t,
\]
where \( P^t \) is \( P \) to the power \( t \). Accordingly, the expected payoff and variance of performance is given by:
\[
\begin{align*}
\hat{V}^t &= x^t \hat{U} = x^0 P^t \hat{U}, \\
\hat{S}^t &= \sum_{m=0}^{N} x^t_m \left( V(m) - \hat{V}^t \right)^2,
\end{align*}
\]
where \( \hat{U} = \begin{bmatrix} V(0) \\ \vdots \\ V(N) \end{bmatrix} \).

The reduced form learning curve can be written as:
\[
V_{LD}^t (u_b, \delta, N) = u_b + d \cdot X_t(N)
\]
where
\[
X_t(N) = x_0 P^t
\]
and \( u_b \) is the lowest possible performance, and \( d \) is the difference, \( u_g - u_b \), between the lowest and highest performance level. This formula describes a learning curve with some of the basic features of observed learning curves. In particular it exhibits the “power law of learning” namely learning initially proceeds quickly and then slows down (Snoddy (1926)). The speed of this effect is determined by the complexity parameter \( N \), with \( u_b \) determining the starting point, and \( d \) determining the maximum gain in performance that is possible.

In this model the matrix \( P \) is upper diagonal, thus the eigenvalues of the matrix are the diagonal elements and are given by \( \lambda_i = i/N, i = \{0, 1, .., N\} \). And hence we may write the formula for mean performance in terms of powers of the eigenvalues, or:\footnote{More precisely we can rewrite \( P = VLM \), where \( L \) is a diagonal matrix with the eigenvalues along the diagonal. Then \( z_i = a_i b_i \), where \( a_i \) and \( b_i \) are the \( i \)th coordinate of the row vector \( x^0 V \) and the column vector \( M \hat{U} \) respectively.}
\[
\hat{V}^t = \sum_{i=0}^{N} z_i t^i \ln \lambda_i.
\]
This is precisely the functional form that Newell and Rosenbloom (1981) suggest provides the best fit data they explore. This implies that a priori this model will provide a good fit to a wide variety of learning curves.

For purposes of comparison consider the following learning curve due to Jovanovic and Nyarko (1995), based upon a simple Bayesian decision problem. In their model it is assumed that the decision maker each period chooses \( z_t \in \mathbb{R} \) to yield performance:

\[
q_t = A \left[ 1 - (y_t - z_t)^2 \right].
\]

The parameter \( y_t \) is not known at the time \( z_t \) is chosen, and is assumed to be normally distributed, satisfying \( y_t = \theta + w_t \), where \( \theta \) is a time invariant constant and \( w \) is a \( N(0, \sigma^2_w) \) random variable, assumed to be serially uncorrelated with time. Though the variance of \( w_t \) is known, the agent does not know \( \theta \), but is assumed to have a normally distributed unbiased estimate \( \hat{\theta} \), with variance \( \sigma^2_\theta \).

At the end of each period the individual observers the realization of \( y_t \), and the optimal strategy each period is to set \( z_t = E(y_t|\hat{\theta}, y_0, y_1, ..., y_{t-1}) \). Thus the expected performance at date \( t \) is given by:

\[
V_{PL}^t(A, \sigma^2_w, \sigma^2_\theta) = A \left[ 1 - \sigma^2_t - \sigma^2_w \right],
\]

where \( \sigma^2_t \) is the posterior variance of \( \theta \) given by:

\[
\sigma^2_t = \frac{\sigma^2_w \sigma^2_\theta}{\sigma^2_w + t \cdot \sigma^2_\theta}.
\]

This model generates a three parameter learning curve, where \( A \) controls the range of learning (between 0 and a maximum of \( A \left[ 1 - \sigma^2_w \right] \)), while \( \sigma^2_\theta \) and \( \sigma^2_w \) jointly determine the speed of learning. Throughout, let index LD refer to the learning by doing model, while PL refers to this Parametric Learning model.

### 3.1 Learning a Corporate Culture

Though the LD and PL models highlight difference aspects of the learning process, they both share the feature of having the same basic properties of a learning curve - performance increases quickly at the beginning, and then slows down with experience. Hence, to empirically distinguish between the models one must depend upon differences in their shapes. This section presents the results from estimating the two learning models using data from an experiment performed by Weber and Camerer (2001). They report results from an experimental study of the effect of mergers on team performance that seem to be appropriate for the learning by doing model introduced in this paper.

---

\(^{10}\)See equations 7 and 8 of Jovanovic and Nyarko (1995).
In their experiment individuals are divided into two groups (firms), in which one person (the manager) is exposed to a sequence of pictures (the events), who then has to communicate this information to the second person (the worker), who executes a task as a function of the picture shown. After several periods of learning, firms were merged by having one manager fired, with the remaining manager required to transmit the information to the remaining two workers.

The loss of performance when mergers occur has two interpretations, depending upon the learning model used. The PL model is fitted to the data by supposing the difference between \( z_t \) and \( y_t \) represents the average error in responses in period \( t \). When a merger occurs, this error is assumed to increase on average, with performance steadily increasing at the same rate. Thus there are 4 parameters in the model, \( A \), variance of the noise parameter, \( \sigma^2_w \), the prior variance of \( \theta \) at the beginning of the experiment and just after the merger, given by \( \sigma^2_{\theta_1} \) and \( \sigma^2_{\theta_2} \) respectively. The learning by doing model supposes that individuals have coded the optimal response for a fraction of the possible states. It is assumed that after merger the number of such states decreases, and then increases with the new manager. This decrease is estimated by \( q \), the probability that the memory for an event is lost, and hence the learning by doing model has 4 parameters, the initial payoff, \( u_b \), the maximum increase in performance possible, \( d \), the number of events, \( N \) and the probability of a forgetting an event after the merger event, \( q \).

Following Jovanovic and Nyarko (1995), the parameters are estimated by minimizing the sum of the squared errors:

\[
\min_{\beta} \sum_{t=1}^{T} (v_t - V_M^t (\hat{\beta}))^2,
\]

where \( v_t \) is observed performance in period \( t \), \( \beta \) is the set of parameters, and \( M \in \{LD, PL\} \) denotes the model. Let \( \hat{\beta}_M \) be the predicted parameter values for model \( M \), \( v = [v_1, ..., v_T]^T \) the vector of data, \( \hat{v}_M = [V_M^1 (\hat{\beta}_M), ..., V_M^T (\hat{\beta}_M)]^T \) the vector of predicted values for model \( M \), and \( J_M = [\nabla V_M^1 (\hat{\beta}_M), ..., \nabla V_M^T (\hat{\beta}_M)]^T \) the Jacobian matrix at the estimated parameter values. There is no Jacobian for \( N \) since it is a discrete variable, and hence is found by systematically searching over the possible values of \( N \), with \( (u_b, d, q) \) estimated at each step to minimize squared error.

In terms of testing we compare the learning by doing model against the parametric learning model using a non-nested hypothesis test, called the \( P \) test, due to Davidson and MacKinnon (1981). The purpose of this test is to see if the alternative hypothesis is to be preferred, and proceeds as follows. Suppose that the null hypothesis is that \( LD \) is correct, then one can write the data as a convex combination of the estimated values from the two models:

\[
v = (1 - \alpha) \hat{v}_{LD} + \alpha \hat{v}_{PL} + \text{residuals}.
\]
Where \( v, \hat{v}_{LD} \) and \( \hat{v}_{PL} \) are vectors with actual performance and estimated performance for each model, each period. The parameter \( \alpha \) is estimated using OLS (ordinary least squares), and determines the optimal combination of estimates from \( LD \) and \( PL \) that fit the data. Davidson and MacKinnon (1981) show that if \( LD \) correct, then the estimated \( \alpha \) should be zero, using the standard errors from the OLS estimate. If not then \( LD \) is rejected. The regression that is run to compute the test is:

\[
v - \hat{v}_{LD} = \hat{J}_{LD} b + \alpha (\hat{v}_{PL} - \hat{v}_{LD}) + \text{residuals},
\]

where \( b \) and \( \alpha \) are both estimated using OLS. This is called a Gauss-Newton artificial regression. The reason that it is artificial is that if \( \hat{J}_{M} \) is estimated precisely then \((v - \hat{v}_{LD}) \hat{J}_{LD} = (0, 0)\), and hence the estimated values of \( b \) should be zero. If not, then the optimization routine needs to have a tighter convergence requirement. Once the regression has been run, one can use the estimated \( t - \text{statistics} \) for \( \alpha \) to test the hypothesis \( \alpha = 0 \). It should be emphasized that rejecting \( LD \), does not imply accepting \( PL \). One has to run a separate regression for model \( PL \), and if both models are rejected when tested against each other then a linear combination of the models provides a better fit, and neither model fits the data better at the exclusion of the other.

The data consists of two series, one with 25 rounds, with merger occurring at round 15, while in the other series there are 30 rounds, with merger occurring in round 20. The results from estimating these models are reported in table 1. These are graphed in figures 1 and 2. Both learning curves do a good job of fitting the data, though the learning by doing model provides a better fit.

### Table 1: Estimates for Corporate Culture Data

<table>
<thead>
<tr>
<th>Data Set</th>
<th>A</th>
<th>( \sigma_{\hat{b}_1}^2 )</th>
<th>( \sigma_{\hat{b}_2}^2 )</th>
<th>( \sigma_w^2 )</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short</td>
<td>899.944</td>
<td>0.753</td>
<td>0.343</td>
<td>0.091</td>
<td>0.956</td>
</tr>
<tr>
<td>Long</td>
<td>856.812</td>
<td>0.735</td>
<td>0.439</td>
<td>0.140</td>
<td>0.910</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Set</th>
<th>( u_b )</th>
<th>( d )</th>
<th>( N )</th>
<th>( q )</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short</td>
<td>-22.241</td>
<td>213.853</td>
<td>5.000</td>
<td>0.241</td>
<td>0.961</td>
</tr>
<tr>
<td>Long</td>
<td>-26.477</td>
<td>237.753</td>
<td>6.000</td>
<td>0.423</td>
<td>0.942</td>
</tr>
</tbody>
</table>

- \( \hat{J}_{LD} \) denotes the optimal combination of estimates from \( LD \) and \( PL \).
Learning by Doing Model  |  Parametric Learning Model  
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>t-statistic</td>
</tr>
<tr>
<td>.070</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Table 2: P-Test Results for Corporate Culture Data

This is particularly evident when the P-test is performed. As one can see in table 2, the LD model is accepted in favor of the PL model, and the PL model is rejected when tested against the LD model. The results highlight the fact that learning in this case is well modelled by a procedure in which individuals learn with experience the optimal responses to a large number of discrete events. It also illustrates a general problem with inference with learning curves, namely they typically all fit well (as the $R^2$ for the PLM of greater than .9 is usually taken as evidence of a good fit). It would have also been the case that if one uses a standard log-linear model, as is typical in the literature (see Alchian (1963)), the fit would also have been good, thought both the PLM and the LD models provide a better fit. These results in and of themselves do not prove that the learning by doing procedure is optimal, but they do illustrate that it is possible to construct an empirically testable model of learning based upon the cost of contingent planning. Additional insight into the model can be found by exploring some of the behavioral implications when the learning by doing algorithm is not optimal.

4 Behavioral Implications

4.1 Probability Matching

It has been shown that when the environment is sufficiently complex and stable, in the sense of proposition 6, then the learning by doing procedure is optimal. This does not imply that it is optimal for all complex environments. In fact, it is well known that when planning is costly, a globally optimal procedure does not exist. In brief, one can use resources to determine an optimal rule, however the determination of how much resources to allocate to such a process is itself a costly optimization exercise, and so one faces an infinite regress of optimization exercises.\(^{11}\)

However, if the learning by doing algorithm is generically a good procedure that is widely adopted by individuals, then this will have implications for observed behavior if used in environments for which it is in fact not optimal. An example of such an environment is the following two armed bandit problem. Each

\(^{11}\)See Day and Pingle (1991) and Lipman (1991) on this issue.
period an individual must select which of two lights, \( L \) or \( R \), will turn on, and receives a fixed reward \( R \) if she guesses correctly.

If the decision maker believes that this is in fact a bandit problem for which the probability of \( L \) is fixed over time, then the optimal strategy is to choose the arm which she believes is most likely to occur, and then to update beliefs over time. In the context of the model of section 2.2 this corresponds to setting \( a_L = \) probability of \( L \). With these parameters \( \Pr(L|R) = \Pr(L|L) = a_L \), while \( \Pr(R|R) = \Pr(R|L) = 1 - a_R \). If the decision maker does not know \( a_L \), then she can use the updating procedure for multinomial random variables described above to form beliefs \( a^t_L \) each period, and choose \( L \) if and only if \( a^t_L \geq 1/2 \). This would result in a behavior in which the decision maker stays with one arm until there are sufficient draws on the other arm to cause her to change her choice. Over time, as beliefs settle down, the individuals will choose one arm only, regardless of the pattern of lights.

What is in fact observed is the well known phenomena of “probability matching” (see Estes (1976) for a literature review), that was widely studied in the 1960’s by psychologists, and considered evidence that people do not make rational decisions, a point that has been much emphasized by Richard Herrnstein (1997). This behavior is described by individuals choosing \( L \) with approximately the probability that \( L \) occurs, rather than sticking to a single choice. From the point of view of Bayesian decision making, such behavior is only irrational if individuals know the true underlying model, which in general was not the case. What is interesting is that the learning by doing algorithm (which is optimal when the process generating the lights is the appropriate Markov chain), predicts precisely this behavior. This is because under the learning by doing algorithm the individual selects the side that was optimal the previous period. Hence the frequency with which \( L \) is chosen is exactly equal to its frequency of occurrence.

This result illustrates the difficulty of testing the “rationality” of behavior using a fixed environment. In particular, the fact that a general “optimization algorithm” does not exist, implies that for any decision making procedure it is possible to find problems for which the observed choices are “irrational”. Hence, individuals should only be judged irrational if they are performing poorly for the decision problems that they are facing on average in practice. If the learning by doing algorithm is on average a good decision procedure, then the fact that individuals engage in probability matching is not necessarily evidence that they are irrational or on average are making poor decisions.

4.2 Similarity Judgements and Optimality in the Long Run

The learning by doing algorithm supposes that individuals improve performance by acquiring experience with events that occur frequently. This raises the issue that in practice an individual is unlikely to experience
**exactly** the same event again in the future. The purpose of this section is to illustrate that the model is robust to allowing individuals to use a rule that associates the same action to events that are “similar”.

There is a voluminous literature in psychology that models individual choice as resulting from the exercise of similarity judgements, which, as Tversky and Kahneman (1981) have shown, implies that in many situation individuals make decisions that are inconsistent with rational choice theory.\(^{12}\) For example, Tversky and Kahneman (1981) show that a subject’s response to the question of whether a vaccine should be given to a population depends upon whether or not the risks are presented in terms of mortality rates or survival rates. These, and many other similar results, illustrate that when responding to a question the typical individuals does not in fact explore all the implications of the data before making a decision. Hence in the context of Newell (1990)’s decision typography, these are choices made in the “cognitive” time frame.

In economics there is a literature, beginning with Luce (1956), that takes similarity judgements as given, and then asks how one may formally model such behavior.\(^{13}\) The question that is not explored in economics is the converse, namely to what extent do similarity judgements lead to rational choice. This question is natural in the context of the learning by doing algorithm when it is extended to deal with more complex event spaces. Suppose that the set of possible events, \(\Omega\), is now a convex subset of \(\mathbb{R}^d\), and that \(\{\omega^t\}_{t=1}^{\infty}\) is an i.i.d. process represented by a measure \(\mu\) on \(\mathbb{R}^d\) that is absolutely continuous with respect to Lebesgue measure (in other words the probability that \(\omega^t = \omega^{t'}\) for \(t \neq t'\) is zero). Let the space of choices be given by \(Z = \{0, 1\}^n\), for some \(n \geq 1\), and that the individual’s utility function, \(U(z|\omega^t)\), bounded and Borel measurable on \(Z \times \Omega\). This ensures the existence of a Borel measurable optimal choice rule \(\sigma^*(\omega^t) \in \arg\max_{z \in Z} U(z|\omega^t)\), but in general the optimal choice may not be continuous function of \(\omega^t\).

Under these assumptions it is the case that the probability of an event occurring again is zero (\(\Pr\{\omega^t = \omega^{t'}\} = 0\) for all \(t \neq t'\)). Therefore, under the learning by doing algorithm as stated in 1, performance can never improve with experience. Thus, a necessary condition for learning based upon experience is that the individual must extrapolate from past experience to decide on how to behave today. An example of such a decision procedure is the nearest neighbor rule. When the individual observes \(\omega^t\), she then finds the event that occurred in the past that is closest to \(\omega^t\), say \(\omega^{t'}\), and then chooses the optimal decision that is associated with the event \(\omega^{t'}\), given by \(\sigma^*\left(\omega^{t'}\right)\), which was previously encoded in the individual’s memory. As in the previous section, it is assumed that after a decision is made the individual has sufficient time to contemplate her

---

\(^{12}\) See Churchland and Sejnowski (1993) for an excellent review of this literature.

\(^{13}\) Rubinstein (1988) and Leland (1998) are papers that build upon the work of Luce (1956) that introduce concepts of similarity into preferences. Gilboa and Schmeidler (1995) present a model in which such judgements are build up from past experiences or cases, while Sonsino (1997), Jehiel (2001) and Samuelson (2000) explore the implications of such judgements for games.
response, and to associate with the experienced event \( \omega^t \) the optimal response \( \sigma^* (\omega^t) \). Hence, the decision \( z^t \) at time \( t \) is the optimal response to the most similar previous event \( \omega^{t'} \), \( z^t = \sigma^* \left( \omega^{t'} \right) \).

More formally, given a history of events and corresponding optimal responses at time \( t \):
\[
H^t = \{ (\omega^1, z^1), (\omega^2, z^2), \ldots, (\omega^{t-1}, z^{t-1}) \},
\]
and given a new event \( \omega^t \), let \( \Omega(\omega^t, k, H^t) \) be the \( k \) closest events, where closeness is measured by the Euclidean distance \( \| \omega' - \omega \| \). Then a generalization of the nearest neighbor rule is given by the \( k \)-nearest neighbor rule.

**Definition 7** The \( k \)-nearest neighbor rule is defined for odd \( k \) by:
\[
\sigma_i \left( \omega^t | H^t \right) = \begin{cases} 
1, & \text{if } \sum_{\omega' \in \Omega(\omega^t, k, H^t)} \sigma^*_i (\omega') / k > 1/2, \\
0, & \text{otherwise.}
\end{cases}
\]

(7)

This rule requires the agent to use the average best response for the \( k \) events most similar (in terms of Euclidean distance) to event \( \omega^t \). Under the optimal rule the expected utility each period of the individual would be \( U^* = E \{ U (\sigma^* (\omega) | \omega) \} = \int_{\omega \in \Omega} U (\sigma^* (\omega) | \omega) \, d\mu \), while the expected utility in period \( t \) would be a function of the history and given by \( U (H^t) = E \{ U (\sigma (\omega | H^t) | \omega) | H^t \} \). The performance of the behavioral rule can be evaluated by looking at performance for a typical history, which is found by taking the expected value over all possible histories, and involves extending \( \mu \) to a measure on \( \Omega^\infty \) in a natural way using the rule \( \sigma (\omega^t | H^t) \). Letting \( U^t = E \{ U (H^t) \} \), one has the following proposition (proved in the appendix).

**Proposition 8** If the optimal decision rule \( \sigma^* (\omega^t) \) is Borel measurable, then under the \( k \)-nearest neighbor rule for any odd \( k = 1, 3, 4 \ldots \) with a Euclidean similarity measure expected performance approaches the optimum:
\[
\lim_{t \to \infty} U^t = U^*
\]

This somewhat surprising result demonstrates the power of using even a very crude similarity measure in ensuring that learning by doing converges to the optimum. It is based upon some rather deep results in mathematics that demonstrates that measurable functions can be approximated by continuous functions. Also the convergence results does not depend upon the individual’s prior believes, and hence illustrates the robustness of the learning by doing procedure described in this paper.

Given that the starting point of the model is the idea that individuals do not have time to determine an optimal response to \( \omega^t \), then one may wonder if the similarity computation itself may also be too difficult to carry out in the time allowed. As it happens, such procedures are very fast, and as Churchland and Sejnowski (1993) observe, the human brain is optimized to carry out such procedures very quickly.
A remarkable feature of the learning by doing procedure augmented with a similarity judgement is that no knowledge of the function $\sigma^*(\omega^t)$ is required. The procedure works with a wide variety of similarity judgements, and illustrates that the learning by doing algorithm is a robust, non-parametric procedure that results in behavior that converges utility maximization in the long run. In particular, the model is consistent with the observation that in certain domains, where the individual has a great deal of experience, performance may be close to optimal, while in domains where experience is limited then behavior may be inefficient. In other words, the model is consistent with the observation that individual performance is very heterogeneous across domains.

The fact that the model predicts that behavior converges to the optimum in the long run is not inconsistent with the recent work in behavioral economics incorporating elements of irrational behavior into economic models. This is because the model makes no prediction regarding behavior in the short run, the case considered by behavioral models. Moreover, one can choose rules $\sigma^*(\omega^t)$ in such a way to make convergence arbitrarily slow. In such cases the utility maximization model would perform quite poorly, and hence the observed behavior may be better predicted using models based upon decision making shortcuts. (see for example the collection of papers in Kahneman, Slovic, and Tversky (1982) reporting early progress on this problem).

5 Concluding Discussion

The goal of this paper is to illustrate how the explicit introduction of planning costs into a model of decision making under uncertainty can result in a theory of learning by doing that is empirically implementable, and consistent with many features of actual decision making. In particular, it can explain why in the short run individuals may make very poor decisions, while over all the hypothesis of utility maximization provides a good first order approximation to observed behavior, particularly after an individual has acquired experience with a particular problem or task. Not only does the model predict that adaptive learning through experience may be an optimal strategy, it can also explain the extreme non-linearity in observed behavior. That is why one simultaneously observes individuals who are capable of carrying out very complex and sophisticated tasks (in domains where they have a great deal of experience), yet make mistakes in elementary decision problems. The learning curves derived from the model procedure are consistent with the empirical evidence in Newell and Rosenbloom (1981), and in the case of the corporate culture data, the learning by doing model
provides a better fit than a more traditional Bayesian learning model.\textsuperscript{14}

The model is not intended to address the central question of behavioral economics, namely the problem of understanding the systematic ways in which individual choices deviate from optimal behavior (see for example Camerer (1995) and Rabin (1998)). Rather, the goal is to better understand the scope of the utility maximization hypothesis. When an individual faces a new event, where for simplicity it has been assumed that either a random choice is made, or a very simple similarity measure is used to guide behavior based upon past experience, then it has been shown that individuals may rationally choose to commit errors, rather than to engage in complete contingent planning. One way to interpret behavioral economics in the context of this model is that its goal is to provide better models for these “errors” when complete planning (i.e. rational choice) is not possible. As Newell (1990) observes, these choices are conscious, and hence amenable to modelling. When there is a stable relationship between events and the associated optimal response, and, moreover, the set of possible events is sufficiently limited that eventually all new events have either been experienced before, or the optimal responses are close to a previously experienced event, then learning by doing results in behavior that converges to rational choice. These observations are consistent with Vernon Smith (1991)’s claim that decision costs may explain many of the behavioral anomalies in experimental market data.

When these conditions are not satisfied, then convergence to optimal behavior can be quite slow. These points are nicely illustrated in the movie \textit{Groundhog Day}, where Phil, played by Bill Murray, must constantly repeat one day in his life in the town of Punxsutawney, Pennsylvania until he finally gets it right. This film not only illustrates how learning by doing eventually leads to optimal behavior (in Phil’s case a meaningful relationship and happiness), but also illustrates how useful it is for Phil to be able to experiment with different strategies for the same set of problems. Unfortunately, the rest of us must simply bumble along, and may only achieve such understanding in the very long run, if ever!

\textbf{References}


\textsuperscript{14}See also the work of Sinclair-Desgagné and Soubeyran (2000) who provide a heuristic model of learning by doing using techniques from optimal control theory.


A Appendix

A.1 Proofs of Propositions

**Proposition** For \( t \geq 0, b > 0, \lambda, \delta \in (0, 1) \) the marginal benefit function \( mb(t, b, \lambda, \delta) \) strictly decreasing in \( t \) and strictly increasing in \( b, \lambda \) and \( \delta \).

**Proof.** Without loss of generality set \( u_g - u_b = 1 \). The existence and continuity of \( mb(t, b, \lambda, \delta) \) follows from the closed form expression for \( \pi(t, t + n, b, \lambda, \delta) \), and equation 2. Observe that \( \pi(t, t + n, b, \lambda) < 1 \), and decreasing with \( n \). Moreover \( mb(t, b, \lambda, 1) = Pr \{ \text{event } \omega \text{ occurs at least once} \} \leq 1 \) and \( \lim_{b \to \infty} mb(t, b, \lambda, 1) = 1 \). From this it immediately follows that marginal benefit is increasing in \( \delta \). This also implies that it is To characterize the other properties of the marginal benefit, notice that it satisfies the following difference equation:

\[
mb(t, b, \lambda, \delta) = \alpha_t + \delta (1 - \alpha_t) mb(t + 1, b, \lambda, \delta),
\]

where \( \alpha_t = \frac{\lambda}{1 + t/b} \).

Hence \( \alpha_t \) is decreasing in \( t \) and is in the interval \((0, 1)\). Thus we can conclude that:

\[
mb(t + 1, b, \lambda, \delta) \leq \frac{\alpha_t}{1 - \delta}.
\]

It follows that:

\[
mb(t, b, \lambda, \delta) - mb(t + 1, b, \lambda, \delta) = \alpha_t + (\delta (1 - \alpha_t) - 1) mb(t + 1, b, \lambda, \delta)
\]

\[
\geq \alpha_t \left( 1 - \frac{(1 - \delta (1 - \alpha_t))}{1 - \delta} \right)
\]

\[
> 0.
\]

Now observe if we differentiate the difference equation one obtains:

\[
\frac{\partial mb(t, b, \lambda, \delta)}{\partial \lambda} = (1 - \delta mb(t + 1, b, \lambda, \delta)) / (1 + t/b) + \delta (1 - \alpha_t) \frac{\partial mb(t + 1, b, \lambda, \delta)}{\partial \lambda}.
\]
Now for $\delta < 1$, $\delta mb(t + 1, b, \lambda, \delta) < 1$, and hence the constant term is positive, and $\delta (1 - \alpha_t) < 1$, from which we conclude that $\partial mb(t, b, \lambda, \delta) / \partial \lambda_t$ exists and is strictly positive. In the case of $b$ one has:

$$\partial mb(t, b, \lambda, \delta) / \partial b = (1 - \delta mb(t + 1, b, \lambda, \delta)) \frac{t \lambda_t}{(b + t)^2} + \delta (1 - \alpha_t) \partial mb(t + 1, b, \lambda, \delta) / \partial b,$$

and hence a similar argument demonstrates that $\partial mb(t, b, \lambda, \delta) / \partial b > 0$. ■

**Proposition** Suppose $\lim_{n \to \infty} mc(n) = \infty$, then for each $b$, there is an $N^*(b)$, such that for $N > N^*(b)$, the optimal level of planning, $n^*(t, b, N)$, satisfies:

$$mc(n^* + 1) \geq mb(t, b, 1/N) \geq mc(n^*). \quad (8)$$

Moreover $n^*(t, b, N)$ has the following properties:

1. $n^*(t, b, N) \geq n^*(t + 1, b, N)$, and there is a $T^*$ such that $n^*(t, b, N) = 0$ for $t \geq T^*$.
2. $n^*(t, b, N)$ is increasing with $b$.
3. If $N > (u_g - u_b) / mc(1)$, then for $b$ sufficiently close to zero $n^*(t, b, N) = 0$ for all $t$.

**Proof.** The hypothesis that $\lim_{n \to \infty} mc(n) = \infty$, ensures the existence of $n^*(t, b, N)$. From statement 1 for each $N$ there is a $T(b, N)$ such that $n^*(t, b, N) = 0$ for $t > T(b, N)$. Moreover, $T(b, N)$ is decreasing with $N$, and hence there exists a smallest $N(b)$ satisfying:

$$N > \sum_{t=0}^{T(b, N)} n^*(t, b, N) + T(b, N).$$

Moreover this expression is satisfied for all $N > N(b)$. The right hand side specifies the largest number of states that one would have explored after $T(b, N)$ periods. For $t > T(b, N)$ only learning by doing is optimal, and hence there is never an option value to delaying the exploration of a state. This ensures that the optimal number of states to be explored each period is $n^*(t, b, N)$, as given by 8.

Let $c = mc(1) > 0$ be the marginal cost of planning for a single state. Notice that

$$\lim_{t \to \infty} \pi(t, t + 1, b, 1/N) = 0,$$

and $\pi(t, t + n, b, 1/N) > \pi(t, t + n + 1, b, 1/N)$, hence it follows that

$$\lim_{t \to \infty} mb(t, b, N) = 0,$$

and for large enough $t$, $mb(t, b, N) < c$, from which statement 1 follows.

For any $T > t$, $\pi(t, T + 1, b, 1/N) = \pi(t, T, b, 1/N) \left( \frac{b(N-1)+T-1}{bN+T} \right)$, from which one can show recursively that $\partial \pi(t, T, b, N) / \partial b > 0$ for $T > t$, and hence the marginal benefit is an increasing function of $b$ from which statement 2 follows.

As $b$ approaches zero this corresponds to the agent placing almost all probability mass on $\Omega_t$ for $t \geq 1$. Thus the only benefit from planning occurs in period 0, before any observations have been made. In that
case the marginal benefit is \((u_g - u_b) / N + \varepsilon (b)\), where \(\lim_{b \to 0}\varepsilon (b) = 0\), hence if \(N > (u_g - u_b) / mc\) (1) the parameter \(b\) can be chosen sufficiently small that no planning takes place, which combined with 1 implies 3.

**Proposition** Suppose planning costs are increasing and convex in \(n\), then there exists an optimal planning rule \(n^0 (r_t, t, b, N)\), where \(r_t\) is the number of states \(\Omega \setminus \Omega^{t-1}\) with the following properties:

1. \(n^0 (r_t, t, b, N)\) is increasing in \(r_t\).
2. If \(r_t \leq \tilde{n} (t, b, N)\), then \(n^0 (r_t, t) = r_t\).
3. If \(r_t > \tilde{n} (t, b, N)\), then \(\min \{ n^* (t, b, N), r_t \} \geq n^0 (r_t, t, b, N) \geq \tilde{n} (t, b, N)\), where \(n^* (t, b, N) = \arg \max_{n \geq 0} n \cdot mb (t, b, N) - c (n)\) is the optimal one period strategy and \(\tilde{n} (t, b, N)\) satisfies:

\[
mb_t - mc (\tilde{n}) \geq \max \{ p_t + \delta (mb_{t+1} - c (1)) , 0 \} \geq mb_t - mc (\tilde{n} + 1),
\]

where \(p_t = (N + t/b)^{-1}\) is the probability that an event that has not been observed occurs in period \(t\), and \(mb_t = mb (t, b, 1/N)\).

**Proof.** Let \(V_t (r_t)\) be the value obtained from further exploration of states, then from the dynamic programming algorithm the optimal search rule, \(n^* (t, b, N)\), and the value function solve:

\[
V_t (r_t) = \max_{n \leq r_t} (n_t \cdot mb_t - c (n_t)) + \delta \left\{ p_t (r_t - n_t) V_{t+1} (r_t - n_t - 1) + (1 - p_t (r_t - n_t)) V_{t+1} (r_t - n_t) \right\}.
\]

The first term is the value of search in the current period, while the second term is the value from delaying search to subsequent periods. Due to the convexity of \(c\) the marginal cost of planning increases with \(n_t\), and hence it maybe worthwhile to delay planning for an event until the next period. The benefit from doing this is at least \(\delta (1 - p_t) (mb_{t+1} - c (1))\), where \((1 - p_t)\) is the probability the event does not occur in period \(t\) (and hence one adds it to the set of known events for period \((t + 1)\), and \(mb_{t+1} - c (1)\) is the benefit for planning this event in the following period. For any \(n_t \leq \tilde{n} (t, b, N)\) it never pays to delay planning, and moreover if \(n_t < \tilde{n} (t, b, N)\), then the agent gains by increasing \(n_t\), hence for \(r_t \leq \tilde{n} (t, b, N)\), \(n^0 (r_t, t, b, N) = r_t\).

Since \(mb_t\) is decreasing with \(t\) and \(\lim_{t \to \infty} mb_t = 0\), then \(\tilde{n} (t, b, N)\) decreases with \(t\). If \(\tilde{n} (0, b, N) < 0\) the individual engages in no planning, and we are done. Suppose not, then there is a first period \(T\) such that \(\tilde{n} (T - 1, b, N) \geq 1\), and \(\tilde{n} (T, b, N) < 0\) from period \(T\) on, and hence it is optimal to do no further search and \(V_t (r_t) = 0\) for all \(t \geq T\), and \(r_t \geq 0\). This implies that the optimal planning problem can be viewed as a finite horizon problem with a finite state space, thus ensuring the existence of an optimal strategy.
The final step is to characterize \( n^0 (r_t, t, b, N) \) and show that it is increasing in \( r_t \) when \( r_t > \hat{n} (t, b, N) \). Let \( \mu_t (r) = V_t (r) - V_t (r - 1) \) be the marginal benefit from having more states to explore. It shall be shown by induction that the function is decreasing with \( r_t \) from which it will follow that the optimal planning strategy is increasing in \( r_t \). From the above it is the case that \( \mu_t (r) = 0 \), and hence the function is decreasing in \( r \) for \( t = T \).

Let \( \lambda_t (n) = mb_t - mc (n) \) be the net marginal benefit in period \( t \) from increasing planning from \( n - 1 \) to \( n \). The first order conditions for the optimal amount of planning satisfy:

\[
\tilde{\mu}_t (r_t - n^0) \geq \lambda_t (n^0) \geq \tilde{\mu}_t (r_t - n^0 + 1),
\]

where

\[
\tilde{\mu}_t (r) = \delta \{ \mu_t (r - 1) \mu_{t+1} (r) + (1 - pr) \mu_{t+1} (r) \}
\]

is the increase in the expected value from increasing the stock of unexplored states from \( r - 1 \) to \( r \). From the induction hypothesis that \( \mu_t (r) \) is decreasing in \( r \) it follows that \( \tilde{\mu}_t (r) \) is also decreasing in \( r \). Since \( \lambda_t (n) \) is also decreasing with \( n \), then it follows from 9 that \( n^0 (r, t, b, N) \) is increasing in \( r \).

Finally, using this result one can show that \( V_t (r) - V_t (r - 1) = \lambda_t (n^0 (r, t, b, N)) \) (there are two cases, depending upon whether or not \( n^0 \) increases when going from \( r - 1 \) to \( r \)), which implies that \( \mu_t (r) \) is decreasing in \( r \), and hence the induction hypothesis is satisfied and we are done. \( \blacksquare \)

Construction of payoff under bandit problem.

**Proposition** Under the \( k \)-nearest neighbor rule with a Euclidean similarity measure expected performance approaches the optimum:

\[
\lim_{t \to \infty} U^t = U^*
\]

**Proof.** Since utility is bounded, then \( M = \max_{z, z', \omega} | U (z|\omega) - U (z'|\omega) | < \infty \). Then we have:

\[
U^* - U^t = E \{ E \{ U (\sigma^* (\omega) | \omega) - U (\sigma (\omega|H^t) | \omega) | H^t \} \} \\
\leq E \{ M \cdot \Pr \{ \sigma^* (\omega) \neq \sigma (\omega|H^t) | H^T \} \} \\
\leq E \left\{ M \cdot \sum_{i=1}^n \Pr \{ \sigma^*_i (\omega) \neq \sigma_i (\omega|H^t) | H^T \} \right\} \\
\leq M \cdot \sum_{i=1}^n \sum_{i=1}^n E \{ \Pr \{ \sigma^*_i (\omega) \neq \sigma_i (\omega|H^t) | H^T \} \}
\]

For a given coordinate \( i \), let \( X_t = \omega^t \) and \( Y_t = \sigma^*_i (\omega^t) \in \{0, 1\} \), then the problem can be viewed as one of prediction, namely given \( X_t \), can one predict \( Y_t \). This is formally the problem studied in the statistical pattern
recognition literature (Devroye, Györfi, and Lugosi (1996)). One measure of the asymptotic performance of the rule is the expected probability of error, or for the $k$–nearest neighbor rule:

$$L_{knn} = \lim_{t \to \infty} E \left\{ \Pr \left\{ \sigma^*_i (\omega) \neq \sigma_i (\omega|H^t) \mid H^T \right\} \right\},$$

which is the limiting value of the last expression in the previous inequality. The best performance occurs if one knows $\sigma^*_i (\omega^t)$, and is given by:

$$L^* = E \left\{ \min \{ \eta (X_i), 1 - \eta (X_i) \} \right\},$$

where $\eta (\omega^t) = \Pr \{ \sigma^*_i (\omega^t) = 1 \}$. Notice that since $\sigma^*_i (\omega^t)$ is a deterministic function, then it is the case that $L^* = 0$. From theorem 5.4 of Devroye, Györfi, and Lugosi (1996) one has that $L^* \leq L_{knn} \leq 2L^*$, and hence one concludes that:

$$\lim_{t \to \infty} M \cdot \sum_{i=1}^{n} E \left\{ \sum_{i=1}^{n} \Pr \left\{ \sigma^*_i (\omega) \neq \sigma_i (\omega|H^t) \mid H^T \right\} \right\} = 0,$$

and we are done. ■
A.2 Figures

Learning and Corporate Culture
Short Series

Figure 1: Learning with Corporate Culture - Short Series
Figure 2: Learning with Corporate Culture Data - Long Series