OPTIMAL LEARNING BY EXPERIMENTATION

N° 9 1 0 4

by

Philippe Aghion*

Patrick Bolton#

Christopher Harris†

Bruno Jullien‡

April 1990

Revised December 1990

We would like to thank Drew Fudenberg, Jérry Green, Andreu Mas–Colell, Eric Maskin, Margaret Meyer, John Moore, Jean–Charles Rochet, Iraj Saniee, and Jean Tirole for helpful comment.. We also benefitted from discussions with seminar participants at Stanford, Harvard, Chicago, UCLA, Cornell, Berkeley, Toulouse, and Paris

---

* DELTA and HEC
# Laboratoire d'Econometrie Ecole Polytechnique
† Nuffield College, Oxford
‡ CEPREMAP

A B S T R A C T

OPTIMAL LEARNING BY EXPERIMENTATION

This paper analyses the dynamic decision problem of an agent who is initially uncertain as to the true shape of his payoff function, but who obtains information about it over time by observing the outcome of his past decisions. In the long run, the action is a short run optimum given the beliefs, but may not be an optimum for the true payoff function. We derive conditions under which the limit action is optimal for the true payoff function and establish the robustness of the results. Finally we study the adjustment process in an example where such complete learning does not achieve in the long run.

Journal of Economic Literature : 020

Keywords : Learning, Experimentation.

R E S U M E

APPRENTISSAGE PAR EXPERIMENTATION

Le papier analyse le problème de choix dynamique d'un individu qui, initialement, ne connait pas sa fonction de gain, mais qui obtient de l'information au cours du temps en observant le résultat de ses décisions antérieures. Dans le long terme, l'action choisie est un optimum de court terme étant données les croyances, mais peut ne pas être optimale pour la vraie fonction de gain. Nous exhibons des conditions sous lesquelles l'action limite est un optimum pour la vraie fonction de gain et établissons la robustesse des résultats. Finalement, nous étudions le processus d'ajustement dans un exemple où l'apprentissage reste incomplet dans le long terme.

Journal of Economic Literature : 020

Mots clef : apprentissage, expérimentation

1 Introduction

This paper analyses the dynamic decision problem of an agent who is initially uncertain as to the true shape of his payoff function, but who obtains information about it over time by observing the outcome of his past decisions. The agent must select an action every period from the same choice set over an infinite number of periods; his decision problem changes over time only to the extent that his information about his true payoff function improves. As long as the agent has not learnt all relevant aspects of his objective function he will be in pursuit of two conflicting objectives: the maximisation of his expected short–run payoff, and the maximisation of the informational content of the current action.[1] We are primarily interested in the limit outcomes of this problem. Under what conditions will the agent's expected short–run payoff converge to his true optimum payoff?

We believe that this question is of importance in many areas of economics. For example, the theory of imperfect competition generally assumes that individual firms know all relevant aspects of the demand function. This assumption is often defended with the argument that if the true demand function is initially unknown, but remains fixed over time, firms eventually learn all relevant aspects of demand from past experience. Thus, if one is primarily interested in the nature of long–run imperfect competition one can usefully simplify the analysis by supposing at the outset that firms know perfectly the demand function they face. A clear statement of this line of argument can already be found in Clower (1959):

> "So long as one deals with a fixed demand function, it is reasonably sensible to suppose that the profit and price calculations of the monopolist are made with reference to this situation in which, following various trial–and–error experiments

---

[1] This trade–off arises in many contexts. See Grossman, Kihlstrom and Mirman (1977) and Kihlstrom, Mirman and Postlewaite (1984) for example.

> with different prices, the monopolist knows the precise character of market demand (at least within some relevant range of price and output quantities)." (Clower (1959) pp 707–708.)

While it is fairly obvious that trial–and–error experiments improve a firm's knowledge about demand, it is much less clear that in the course of optimal experimentation the firm ends up knowing the exact shape of market demand. For experimentation is costly, and optimal learning may dictate that experimentation be stopped before all relevant aspects of demand are known. In fact there exist several examples in the literature demonstrating the possibility that optimal experimentation may not result in adequate learning, most notably Rothschild (1974), McLennan (1984), and Easley and Kiefer (1988). (Adequate learning occurs when, with probability one, the agent acquires enough information to allow him to obtain the true maximum payoff.) On the other hand, it is not too difficult to construct plausible examples where optimal experimentation does result in adequate learning.

Our paper is a first attempt at characterising those situations where adequate learning obtains and those where it does not. We suggest a two–stage approach to the problem of determining under what conditions optimal experimentation leads to adequate learning: first understand the case where the agent's payoff function is deterministic, so that the agent's inference problem is not complicated by the presence of noise; then extend this understanding to take into account the additional issues that arise when noise is present.

In this paper we concentrate primarily on the first step of this approach. On the positive side we show that adequate learning obtains if:

(a) the payoff functior is analytic;

(b) the payoff function is smooth and quasiconcave;

(c) there is no discounting.

It is worth pointing out the intuition behind cases (a) and (b): in each of these cases the

agent can learn how to obtain the true maximum payoff from information gathered by local experimentation. Such experimentation gives him an arbitrarily precise estimate of the slope of the true payoff function at any given point, in case (b), so that he learns, roughly speaking, in which direction he should change his action in order to increase his short–run payoff. Eventually he converges to a point where the estimate of the slope is zero; at this point he obtains the true maximum payoff. Similarly, in case (a), local experimentation provides arbitrarily precise global information about the payoff function so that the agent eventually learns where the maximum payoff is located by incurring arbitrarily small experimentation costs.

On the negative side, we give examples to show that inadequate learning may obtain when the payoff function is:

(a) is smooth but not analytic;

(b) smooth but not quasiconcave;

(c) quasiconcave but not smooth.

(Inadequate learning occurs when, with probability one, the agent fails to acquire enough information to allow him to obtain the true maximum payoff.) Inadequate learning may obtain because local experimentation either does not provide all relevant information (cases (a) and (b)) or does not provide enough information to compensate for the costs involved (case (c)). As Alchian (1950) puts it, case (b) can be understood with the help of the following analogy:

> "A nearsighted grasshopper on a mound of rocks can crawl to the top of a particular rock. But there is no assurance that he can also get to the top of the mound, for he might have to descend for a while or hop to new rocks." (Alchian (1950) p. 31.)

The possibility of both adequate and inadequate learning leads to the question of which is more likely. One way of posing the question more precisely at the theoretical level is to ask what the generic outcome is. We argue that, in the deterministic problem,

genericity is most naturally formulated in terms of the agent's priors. We further argue that, when genericity is formulated in this way, both adequate and inadequate learning are non–generic.

When adequate learning does not obtain one cannot understand the long–run outcome independently of the priors or the adjustment process by which it was reached. An entirely new kind of analysis is called for in these cases: in order to determine the nature of the long–run behaviour of the agent, one needs to characterise the optimal learning strategy. Unfortunately this is possible analytically only in very simple learning problems. Section 6 illustrates the kind of issues arising and the kind of analysis needed, when inadequate learning obtains, in a simple learning problem.[2]

A few general results concerning the long–run behaviour of our problem are central to our analysis. Easley and Kiefer derive such results, but their analysis, unfortunately, specifically excludes the deterministic case. We therefore need to generalize their work by providing a unified framework covering both the deterministic and stochastic case.

The deterministic case does, however, have two potentially troubling features. First, existence may fail. Secondly, by making a smaller and smaller experiment one can simultaneously reduce costs and improve information; whereas in the stochastic case one would, on the contrary, expect the informational content of an experiment to be smaller, the smaller the deviation from the status quo. However, we show that the introduction of even small amounts of noise eliminates the existence problem. Moreover our adequate–learning results are robust to the existence of such noise. These findings suggest that both features stem from a single cause, namely a minor closure problem.

---

[2] The possibility of progress here derives, in part, from the fact that the payoff function is deterministic. The advantages of such payoff functions have been exploited fruitfully in other contexts by Alpern and Snower (1987a, 1987b, 1988), Reyniers (1989a, 1989b) and Rob (1988). They also underline the relevance of our asymptotic analysis of the deterministic case.

The paper is organised as follows: Section 2 sets out the model and formulates the learning problem in such a way that there is a clear separation between the information–gathering and payoff–accumulation aspects of the agent's decision in each period. This formulation highlights the fundamental trade–off between the conflicting objectives of learning and obtaining high current payoffs.[3] It also turns out to be more natural mathematically, leading to a streamlined set of regularity conditions, the role of which in the analysis is, we hope, transparent. All of the general results about the long–run behaviour of our problem are stated in this section. Section 3 contains all the adequate learning results. Section 4 addresses the issues of existence and robustness. Aside from the obvious requirements that the set of actions available to the agent be compact, and that his payoff function be continuous in an appropriate sense, there is only one condition required for existence: the agent's observations must be (at least slightly) noisy. (Mathematically, this can be expressed by saying that the distribution of the agent's observations varies norm continuously with his action and the unknown parameter.) In particular, the familiar common–support assumption is not needed. Section 5 provides an example where adequate learning does not obtain when the payoff function is smooth but not quasiconcave. Section 6 examines an example of inadequate learning. This example shows, among other things, that experimentation may cease altogether after a finite number of periods, even though adequate information has not been acquired; and that experimentation may continue forever, yet adequate information may not be acquired, not even asymptotically. Section 7 provides concluding comments.

---

[3] In our framework, both the payoff obtained in a given period and the signal observed depend on the action taken, the underlying parameter of interest, and a shock. So the optimising agent faces a simple trade–off between choosing his action to maximise his payoff, and choosing it to obtain the best possible signal. In Easley and Kiefer's framework, the signal depends on the action, the parameter, and a shock, but the payoff depends on the action and the signal. So the agent must consider the direct effect of his action on his signal, its direct effect on his payoff, and its indirect effect (via the signal) on his payoff.

2 The Model

The general decision problem can be described as follows. At the outset Nature chooses $\theta \in \Theta$, a parameter describing the environment of the decision–making agent. This choice is not observed by the agent. In each subsequent period t ($t = 1,2,...$), the agent chooses an action $x_t \in X$. A shock $z_t \in Z$ is then realized This leads to a signal $y_t = a(x_t,\theta,z_t) \in Y$ and a payoff $\pi_t = b(x_t,\theta,z_t)$. The agent observes y_t but not π_t. (Note that this formulation does not by any means rule out the possibility that the agent observes his payoff. Indeed, he will observe his payoff whenever it is included in the vector of signals y_t.) The agent's overall objective is to maximise the expected net present value $E[(1-\delta)\Sigma_{t=0}^{\infty} \delta^t \pi_t]$, where $0 \leq \delta < 1$ is the discount factor. Thus, the agent's problem is to choose a sequence of strategies $s = (s_1,...,s_t,..)$ to maximise this expectation, where the strategy in period $t+1$ specifies the action x_{t+1} as a function of the past observations up to $t+1$: $x_{t+1} = s_{t+1}(y_1,...,y_t)$.

We make the following assumptions on the data of our model, which will remain in force throughout this section:

(A1) X, Y, Z and Θ are complete separable metric spaces;

(A2) a is Borel measurable, and continuous in (x,θ);

(A3) b is bounded, Borel measurable, and continuous in (x,θ).

It will be seen at once that (A1), and the requirement that a and b be Borel measurable, are purely technical. The boundedness of b simply ensures that the agent's objective is well defined for all s.[4] The role of the continuity of a and b is more subtle, but the basic idea is clear enough: if the agent's observations and payoff are continuous, then we

---

[4] (A3) can be relaxed somewhat.

can prove continuity results linking the agent's behaviour for large t with his behaviour in a certain long–run, or asymptotic, problem. (A more detailed explanation of the role played by the continuity of a and b is given in the Appendix.)

The sequence of shocks $\{z_t | t \geq 1\}$ is taken to be i.i.d. with distribution R. The parameter θ is distributed independently of the z_t, and has distribution Q. Thus a complete description of the underlying state of the world ω takes the form $(\theta, z_1, .., z_t, ..)$, the set of states of the world is $\Omega = \Theta \times Z^{\infty}$, and the agent's prior over Ω is just the probability measure $P = Q \otimes R^{\infty}$.

Suppose that the agent employs a strategy s, which may or may not be optimal. In period one the agent has no information about the state of the world other than his prior P. In period two, however, he will have made one observation y_1 which, together with his knowledge of his strategy s_1, allows him to revise this prior. Let $P_1(\cdot | \omega)$ be the agent's posterior about the state of the world based on the information available to him in period two when the true state of the world is ω. This posterior will, in general, incorporate information about both θ and z_1, but no information about z_t for $t \geq 2$. More interesting, therefore, is the agent's posterior about the parameter of interest θ, which is the marginal of $P_1(\cdot | \omega)$ over Θ. Denote this by $Q_1(\cdot | \omega)$. More generally, denote by $P_t(\cdot | \omega)$ and $Q_t(\cdot | \omega)$ the agent's posteriors about ω and θ respectively, given his strategy s and the observations $(y_1, .., y_t)$. Finally, denote by $P_{\infty}(\cdot | \omega)$ and $Q_{\infty}(\cdot | \omega)$ the hypothetical posteriors that the agent would have if he could observe the entire sequence $(y_1, ..., y_t, ...)$. (Note that P_t and Q_t $(1 \leq t \leq \infty)$ are random variables whose values are probability measures.)

It is easy to see that the posteriors Q_t must follow a martingale. For clearly the agent's best guess, in period t, as to his posterior in any later period is simply his posterior in period t. This implies, by the martingale convergence theorem, that Q_t will converge, as $t \to \infty$, to a limiting posterior. That is, there is a limit to the information that can be obtained by following strategy s. Indeed, the martingale convergence theorem even implies

that this limit is Q_∞. That is, in the limit the agent obtains all the information he could conceivably obtain, namely that conveyed by observation of the entire sequence $(y_1,\ldots,y_t,\ldots)$. We summarise these remarks in the following observation.

Theorem 2.1 The posteriors Q_t follow a martingale. Moreover $Q_t \to Q_\infty$ with probability one as $t \to \infty$. □

We turn now to the problem of choosing an optimal strategy, and of finding the properties of such a strategy. It is convenient to begin with a simplified version of our model in which there is no learning. More precisely, we consider the version of our model in which the agent does not observe the y_t. It is convenient to regard this model as being parameterised by the agent's prior Q over Θ. Since the agent never acquires any information in this model, his problem is completely stationary. In order to solve it, it is therefore sufficient for him to find an action that maximises $E[b(x,\theta,z_1)]$, and to repeat that action forever. Actually, under our present assumptions, he may not be able to maximise $E[b(x,\theta,z_1)]$.[5] But he will be able to approach the payoff

$$m(Q) = \sup_x E[b(x,\theta,z_1)] \qquad \ldots(2.1)$$

arbitrarily closely.

The original model, too, can be parameterised by Q. Let

$$v(Q) = \sup_s E[(1-\delta)\sum_{t=1}^{\infty} \delta^{t-1}\pi_t] \qquad \ldots(2.2).$$

Once again, our present assumptions are not sufficient to ensure that this payoff can be

---

5 We have not assumed that X is compact.

attained. However, at this stage we are primarily interested in characterising optima for our model, and this problem is logically distinct from that of proving that it has an optimum. We shall have to tackle the question of existence later.

One obvious property of m and v is that $v(Q) \geq m(Q)$ for all Q – the agent can do at least as well when learning is possible as he can when it is not. Another is that m and v are both convex. This means that, if the prior probability is Q with probability λ and Q' with probability $1 - \lambda$, then the agent can do at least as well when he is told which before choosing his strategy as he can when he is not. In other words, more information is once again a good thing. A less obvious property, but one which is fundamental to much of the discussion that follows, is that m and v are both lower semicontinuous.[6]

Note that the problem (2.2) is stationary in the sense that the agent's decision-problem in period t is the same as the problem in period zero, except that the prior Q must be replaced by the posterior $Q_t \equiv Q_t(\cdot \mid \omega)$. The value function $v(\cdot)$ is therefore the solution of the Bellman equation:

$$v(Q) = \sup_x (1 - \delta)E[b(x,\theta,z_1)] + \delta\, E[v(q(Q,x,y_1))] \tag{2.3}$$

where $q(Q,x,y_1)$ is the posterior given y_1 and x. The first term on the RHS represents the expected one-period payoff and the second term represents the continuation payoff which incorporates the value of information obtained from one observation, y_1. If there was no information whatsoever to be obtained from the observation of y_1, then (2.3) would reduce to finding x to maximize $E[b(x,\theta,z_1)]$. If the short–run payoff were independent of x, then (2.3) would reduce to finding x to maximize $E[v(q(Q,x,y_1))]$. Thus the Bellman equation is a rather intuitive representation of the trade–off underlying our decision problem, the trade-off between the maximisation of short–run payoffs and the

---

6 The lower semicontinuity of m follows from that of b. The lower semicontinuity of v follows from the lower semicontinuity of b and the continuity of a.

maximisation of the informational content of x. In general, the action x which maximises short–run payoffs will not be the same as the action that maximises informational content. When this conflict arises, the agent will have to sacrifice part of his short–run payoff in order to acquire useful information for the future. Most situations where there is learning by experimentation are characterised by this fundamental trade–off.

We have already pointed out that $v(Q) \geq m(Q)$. So $v(Q) - m(Q)$ can be regarded as the value of the opportunity to learn in the dynamic model. Our second result concerns the asymptotic behaviour of this quantity.

Theorem 2.2 Suppose that s is optimal. Then $v(Q_t) - m(Q_t) \to 0$ a.s.

Proof See Appendix. □

That is, at the optimum, all valuable learning opportunities are exhausted in the long run. This is not the same thing as saying that, in the long run, the agent learns everything there is to learn. After all, learning is costly.

Now just as Q can be thought of as summarising all information relevant to the original problem, so Q_t can be thought of as summarising the problem faced by the agent in period $t + 1$. Moreover, since $Q_t \to Q_\infty$, we know that the problem at $t + 1$ settles down to some asymptotic problem as t gets large. Let us examine the relationship between the problem for large t and the asymptotic problem more closely.

We begin by observing that $v(Q_t) \to v(Q_\infty)$ and $m(Q_t) \to m(Q_\infty)$. These results follow from the facts that m and v are convex and lower semicontinuous, and that Q_t follows a martingale. Combined with Theorem 2.2, they imply Theorem 2.3.

Theorem 2.3 Suppose that s is optimal. Then $v(Q_\infty) = m(Q_\infty)$ a.s.

Proof See Appendix. □

In other words, in the asymptotic problem the agent knows so much that he cannot derive any benefit from trying to learn more. Once again, this is not the same thing as saying that the agent knows everything there is to know. It is, nonetheless, a very strong result. In particular, under certain circumstances it implies that the agent knows everything that is worth knowing.

We can also obtain a continuity result about the actions chosen by the agent. To state this result precisely, we need some notation. Let $B(x,Q) = E[b(x,\theta,z)]$ for all x. Then $B(\cdot,Q)$ is the agent's best guess of his short-run payoff function when his prior is Q.

Theorem 2.4 Suppose that s is optimal. Then with probability one, all limit points of $\{x_t \mid 1 \leq t < \infty\}$ maximise $B(\cdot,Q_\infty)$.[7]

Proof See Appendix. □

At its simplest level, Theorem 2.4 tells us that, as time goes on, the agent's actions come closer and closer to maximising his best guess of his short–run expected payoff. The intuition behind this result is as follows. As time goes on, the cost of learning, measured in terms of the fall in short–run payoff, remains constant. The benefit of learning, by contrast, falls. Hence the motive of maximising his short–run payoff comes to predominate in the agent's choice of action, and the motive of learning induces smaller and smaller deviations from the set of actions that maximises his short–run payoff. At a more general level, Theorem 2.4 tells us that any limiting action solves the asymptotic problem.[8]

---

7 This result uses the upper semicontinuity of b.

8 To see this, simply combine Theorem 2.4, which tells that any limiting action solves the asymptotic problem without learning, with Theorem 2.3, which tells us that solving the asymptotic problem without learning is sufficient to solve the asymptotic problem with

In order to state Theorem 2.5 we will need some terminology. For any given x and θ, $a(x,\theta,\cdot)$ can be regarded as a random variable on Z, Z being given the probability measure R. We refer to the distribution of this random variable as the observation distribution associated with action x and parameter θ.

Theorem 2.5 With probability one, the agent learns the observation distribution associated with every limit point of $\{x_t \mid 1 \leq t < \infty\}$ and the true parameter. □

We do not prove this theorem, since it is not central to our paper. It is, however, easy to see why it must be true. Indeed, according to one version of the law of large numbers, the empirical distribution obtained by making k independent drawings from a given population distribution converges to the population distribution as $k \to \infty$ with probability one.[9] Now suppose that k independent drawings are made from k possibly different population distributions, and that the population distributions converge to a limiting distribution. Then the empirical distribution obtained will still converge, this time to the limiting population distribution. For ultimately all the drawings are made from distributions that are essentially the same as the limiting population distribution. And this is precisely the situation that arises in Theorem 2.5.

The analysis so far can be summarised in two points. First, the problem for large t is closely related to the asymptotic problem.[10] Secondly, a very strong result holds for the asymptotic problem: $v(Q_\infty) = m(Q_\infty)$. That is, in the asymptotic problem, the agent knows so much that he cannot derive any benefit from trying to learn more. Since learning is costly, this does not automatically imply that he knows everything there is to know.

---

learning.

9 This law is usually referred to as the Glivenko–Cantelli lemma, see Parthasarathy [1967; section II.7].

10 We have continuity results in values, actions, and observations.

But under certain circumstances it does imply that he knows everything that is worth knowing.

The reader will recall that our model has incomplete information and learning, and also that we have already considered the variation on it in which there is no learning. In order to make the ideas of the preceding paragraph precise, it is helpful to consider a second variation on our model, in which the agent is informed of the parameter θ. In this complete-information version of our model, the best payoff the agent can achieve is $E(M(\theta))$, where $M(\theta) = \sup_x B(x,\theta)$ and, in a convenient abuse of notation, $B(x,\theta)$ is $B(x,Q)$ in the case where Q is concentrated at the single point $\theta \in \Theta$. We refer to this payoff as the complete–information payoff.

Next, if Q is concentrated at a single point θ, then we say that the agent has complete knowledge. This corresponds to knowing everything there is to know. Also, if $m(Q) = E(M(\theta))$, then we say that the agent has adequate knowledge. This corresponds to knowing everything that is worth knowing. It is strictly weaker than complete knowledge; for, even if the agent does not know the parameter exactly, he may still know of an action that is optimal irrespective of the parameter values; or, to put the point a different way, he may be able to achieve the complete–information payoff even when he does not have complete knowledge of the parameter.

Finally suppose that $Q_t(\cdot \mid \omega) \to \delta_\theta$ for some state of the world ω. (That is, the limiting beliefs are concentrated at the single point θ.) Then, asymptotically, the agent acquires complete knowledge when ω is the state of the world. If $Q_t(\cdot \mid \omega) \to \delta_\theta$ with probability one, i.e. if the agent acquires complete knowledge with probability one, then we say that complete learning occurs. Similarly, if $m(Q_t(\cdot \mid \omega)) \to M(\theta)$ for a particular ω, then the agent acquires adequate knowledge asymptotically. If $m(Q_t(\cdot \mid \omega)) \to M(\theta)$ with probability one, we say that adequate learning occurs. Since certainly $m(Q_t(\cdot \mid \omega)) \leq M(\theta)$ for all θ, a necessary and sufficient condition for adequate learning is that $E[m(Q_t(\cdot \mid \omega))] \to E[M(\theta)]$.

The import of the conclusion $m(Q_\infty) = v(Q_\infty)$ a.s. is therefore this. In trying to prove adequate–learning results, we can try to show that, for all Q, $m(Q) = v(Q)$ implies that $m(Q) = \int M(\theta)dQ(\theta)$. That is, the only way it can happen that there are no learning possibilities whatever is if the agent is actually maximising the true payoff function with probability one. For then $m(Q_\infty) = v(Q_\infty)$ a.s. implies that $m(\theta_\infty(\cdot\,|\,\omega))$ $= \int M(\varphi)d\theta_\infty(\varphi|\,\omega)$ for almost all ω. Moreover our continuity results show that $m(\theta_t(\cdot\,|\,\omega))$ $\to m(Q_\infty(\cdot\,|\,\omega))$ a.s. Hence, overall, $E[m(Q_t(\cdot\,|\,\omega))] \to E[m(Q_\infty(\cdot\,|\,\omega)] = E[\int M(\varphi)dQ_\infty(\varphi|\,\omega)]$, and the latter is just $E[M(\theta)]$ by Fubini's theorem. So adequate learning does obtain.

This is precisely the way in which we shall set about proving adequate learning results in Section 3. However, knowing that such results can be obtained, the following result should already be of interest.

Theorem 2.6 Suppose that s is optimal. Suppose too that adequate learning occurs. Then, with probability one, every limit point of $\{x_t\,|\,1 \leq t < \infty\}$ maximises $B(\cdot,\theta)$.

Proof See Appendix □

Theorem 2.6 is closely related to Theorem 2.4. It shows that if the conditions of Theorem 2.4 are strengthened by requiring that adequate learning occurs, then a correspondingly stronger result is obtained – every limiting action maximises the true short-run payoff function.

3 Adequate Learning Results

In this section we describe three cases in which adequate learning occurs. The three cases all share one common feature: we assume that $y_t = \pi_t$. That is, each period the agent is informed of his payoff, and this is how he learns. In the first case it is assumed further that the payoff function is real analytic with probability one, and that there is no noise. In the second it is assumed the payoff function is continuously differentiable and quasiconcave, and again that there is no noise. In the third case noise is allowed, but we assume that there is no discounting.

3.1 The analytic case

Consider first the case in which the payoff function is a polynomial and there is no noise. More precisely, suppose that

$$b(x,\theta,z) = \sum_{i=0}^{I} c_i(\theta)x^i.$$

Then the agent will acquire adequate knowledge after at most $I + 1$ periods. To see why, suppose that s is an optimal strategy. If s involves $I + 1$ distinct actions over the first $I + 1$ periods, then the agent will be able to calculate all the coefficients of the polynomial at the end of period $I + 1$, and will therefore know the global shape of his payoff function. If, on the other hand, s involves fewer actions, then the agent can change some of them very slightly. Because his payoff function is continuous, this will hardly affect his payoff from the first $I + 1$ periods. And it will ensure that he knows the global shape of his payoff function at the end of period $I + 1$. So either way the agent can be sure of obtaining the complete–information payoff from period $I + 2$ onwards.

More formally, we introduce the following temporary assumptions:

(T1) $a \equiv b$;

(T2) b depends only on x and θ;

(T3) $X = [\underline{x},\overline{x}] \subset \mathbb{R}$;

(T4) $b(\cdot,\theta)$ is real analytic on X with probability one.

(T1) simply states that the agent learns his payoff; (T2) eliminates noise; (T3) is self-explanatory; and (T4) means that, for any $x \in X$, $b(\cdot,\theta)$ can be expanded in a power series about x in a neighbourhood containing x.

Theorem 3.1 Suppose that (A1)–(A3) and (T1)–(T4) hold. Then any optimal strategy s involves adequate learning. □

Notice that this result generalises the result for polynomials in two ways. It allows for an infinite number of coefficients, and it allows for $b(\cdot,\theta)$ that can be expanded locally but not globally.

It is not obvious that the result must hold. For while it is certainly true that the agent can learn all the coefficients at essentially no cost, the benefit from doing so is only obtained after an infinite number of periods!

The essence of the proof is this. We know that $v(Q_\infty) = m(Q_\infty)$. Hence it suffices to show that, if the original problem is such that $v(Q) = m(Q)$, then the agent has adequate knowledge in that problem. To this end, suppose that $m(Q) = v(Q)$ but $m(Q) < E(M(\theta))$. Then one optimal strategy for the agent is to find x^* to maximise his best guess $B(\cdot,Q)$ of his payoff function, and choose x^* forever. He can, however, improve on his payoff from this strategy, which is a contradiction. All he need do is experiment over a small neighbourhood of x^* for n periods. In this way he obtains an approximation to the first n

derivatives of $b(\cdot,\theta)$. Since $b(\cdot,\theta)$ is analytic, the approximate derivatives at a single point can be used to arrive at an approximate picture of the global behaviour of $b(\cdot,\theta)$. He can therefore pick the action that is optimal according to this picture. This procedure works because choosing a smaller and smaller neighbourhood around x^* for the purposes of experimentation simultaneously reduces the cost of experimentation and increases the accuracy of the estimates of the derivatives of $b(\cdot,\theta)$. The formal proof is given in the Appendix.

Corollary Suppose that the conditions of Theorem 3.1 hold. Then the agent's payoff in period t, namely $b(x_t,\theta)$, converges to $M(\theta)$ with probability one.

In particular, every limiting action maximises the true payoff function.

Proof This follows from Theorem 2.6 and Theorem 3.1. □

Note finally that Theorem 3.1 can certainly be extended to the case of a real-analytic function of several variables.[11]

3.2 The smooth quasiconcave case

We have already seen that, if the payoff function $b(\cdot,\theta)$ is continuous, the cost of local experimentation is very small. One might therefore conjecture that the agent will continue experimenting as long as he knows that there is some chance that he is not at a local maximum, and that the sequence of actions chosen will converge to a local maximum. This conjecture isn't quite correct as it stands. To see why, suppose that the agent's current action is x and that he is at a local maximum with high probability. If he tries action $x + \epsilon$ then, with high probability, he incurs a small loss. In this case he will return

---

11 X would need to be in some sense connected.

to action x. On the other hand, there is a small probability that he obtains a small gain. In this case he can continue with action $x + \epsilon$, thereby capitalising on the gain. But his overall expected gain may nonetheless be negative, because of discounting.

This example highlights the essential problem. It is that although the costs of local experimentation can be made small, the information obtained only leads to local improvements, which are likewise small. This problem did not arise in the analytic case, in which local experimentation led to global improvements. A version of the conjecture can nonetheless be salvaged. Indeed, suppose that $b(\cdot,\theta)$ is differentiable rather than continuous, and that the agent's current action is x. If he tries action $x + \epsilon$ then one of two things may happen. First, it may turn out that $\partial b/\partial x(x,\theta) < 0$. In this case he can switch to action $x - \alpha\epsilon$, for as long as he wishes. Secondly, it may turn out that $\partial b/\partial x(x,\theta) > 0$. In this case he can switch to action $x + \alpha\epsilon$ for as long as he wishes. His gain will therefore be at least $\epsilon\partial b/\partial x(x,\theta) + \delta\alpha\epsilon|\partial b/\partial x(x,\theta)|$. Hence, as long as there is any chance that $\partial b/\partial x \neq 0$ at x, he can ensure that his overall expected gain is positive by choosing ϵ sufficiently small and α sufficiently large.

The crucial difference between continuity and differentiability is this. With continuity the agent can discover, at essentially no cost, the direction in which he should move. But he does not know how far to move. Indeed, one can even construct an example in which the agent knows that his payoff function is strictly increasing with probability one, but cannot take advantage of this fact because he does not know how far he should move to the right. With differentiability, on the other hand, this difficulty can be avoided.

We formalize our conjecture using the following assumptions:

(T1) $a \equiv b$;

(T2) b depends only on x and θ;

(T3) $X = [\underline{x},\overline{x}] \subset \mathbb{R}$;

(T4) for almost all θ, $b(\cdot,\theta)$ is continuously differentiable and quasiconcave;

(T5) the function $D(\theta) = \max_x |\partial b(x,\theta)/\partial x|$ is integrable.

As in Section 3.1, these assumptions are temporary, and they are additional to (A1)–(A4). The differentiability part of (T4) ensures, roughly speaking, that experimentation cannot stop before a stationary point is reached, and the quasiconcavity ensures that any stationary point is also a global maximum. (T5) is a technical assumption. It ensures that we can differentiate under the expectation sign.

Theorem 3.2 Suppose that (A1)–(A3) and (T1)–(T5) hold. Then any optimal strategy s involves adequate learning.

Proof By (T5) we know that $\int D dQ < \infty$. Since $D \geq 0$, it follows that $\int D dQ_\infty < \infty$ a.s. Hence, as in the proof of Theorem 3.1, it suffices to show that $m(Q) = v(Q)$ implies that the agent has adequate knowledge. Let x^* maximise $B(\cdot,\theta)$. In view of quasiconcavity, it suffices to show that if $x^* < \bar{x}$ then $\partial b/\partial x(x^*,\theta) \leq 0$ with probability one, and that if $x^* > \underline{x}$ then $\partial b/\partial x(x^*,\theta) \geq 0$ with probability one. Let us treat the case $x^* < \bar{x}$.

Suppose that $x^* + \alpha\epsilon < \bar{x}$. then the following deviation from s is feasible: (i) play x^* in period 1; (ii) play $x^* + \epsilon$ in period 2; (iii) if $b(x^* + \epsilon,\theta) \leq b(x^*,\theta)$ then play x^* forever more; (iv) if $b(x^* + \epsilon,\theta) > b(x^*,\theta)$ then play $x^* + \alpha\epsilon$ forever more. It leads to an increase in the agent's payoff of

$$\delta(1-\delta)E\Big[b(x^* + \epsilon,\theta) - b(x^*,\theta)\Big] + \delta^2 E\Big[(b(x^* + \alpha\epsilon,\theta) - b(x^*,\theta))\chi(b(x^* + \epsilon,\theta) - b(x^*,\theta))\Big],$$

where χ is the indicator function of $(0,\infty)$. Since playing x^* forever is optimal, this increase is non–positive. We may therefore divide by ϵ and let $\epsilon \to 0$ to conclude that

$$0 \geq (1-\delta)E\left[\frac{\partial b}{\partial x}(x^*,\theta)\right] + \delta\alpha E\left[\max\{0,\frac{\partial b}{\partial x}(x^*,\theta)\}\right].$$

Since we are free to choose α as large as we like, this implies that $\partial b/\partial x(x^*,\theta) \leq 0$ with probability one, as required. □

Corollary Suppose that the conditions of Theorem 3.2 hold. Then the agent's payoff in period t, namely $b(x_t,\theta)$, converges to $M(\theta)$ with probability one.

Proof This follows from Theorem 2.6 and Theorem 3.2. □

Theorem 3.2 and its corollary can be extended in various ways. One extension is to the case of many dimensions.[12] Another involves abandoning quasiconcavity. If this is done then one can prove that, with probability one, every limit point of $\{x_t\}$ satisfies the first-order necessary conditions for an optimum. If, moreover, $b(\cdot,\theta)$ is twice continuously differentiable, then every limit point satisfies the second order conditions too. Hence, if it is never the case that both $\partial b/\partial x(x,\theta)$ and $\partial^2 b/\partial x^2(x,\theta)$ are simultaneously zero, then every limit point is a local maximum.

3.3 The undiscounted case

In Section 3.1 and 3.2 we excluded noise from our model, and showed how this permitted effective learning strategies on the part of the agent. These strategies did, however, depend on regularity properties for the payoff function. In this subsection we consider a different possibility: if the discount rate is low then the costs of experimentation will be low compared with the potential benefits, so a great deal of learning can be

---

[12] Here it can be shown that if (T3) is replaced by the requirement that X is a convex subset of $\mathbb{R}^N$, and if (T5) is replaced by the assumption that $\sup|\nabla b(\cdot,\theta)|$ is integrable, then Theorem 3.2 continues to hold. (Compactness of X is not required since existence is assumed.)

expected. More precisely, we show that if there is no discounting, then, even if there is noise, the complete–information payoff is attained under very general conditions on the payoff function. This highlights the difference between noise and discounting. Noise makes learning more difficult, whereas discounting makes learning less attractive.

We must extend our existing definition of the agent's payoff to cover the case $\delta = 1$. We take it to be

$$E\left[\liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \pi_t\right].$$

This is the most conservative definition of the payoff that is possible. (For example, $\liminf E\left[\frac{1}{T}\Sigma_{t=1}^{T}\pi_t\right] \geq E\left[\liminf \frac{1}{T}\Sigma_{t=1}^{T}\pi_t\right]$ by Fatou's lemma.) We shall nonetheless show that, even on this definition, the payoff can be made to attain the complete–information level.

We make the following assumptions:

(T1) $a \equiv b$;

(T2) b is bounded and Borel measurable; moreover it is lower semicontinuous in x.

In order to understand (T2), note that there are a countable number of periods only, so it is almost a necessary condition for adequate learning that everything worth knowing about the short–run payoff functions $B(\cdot \mid \theta)$ can be learnt from a countable number of experiments. For example, suppose that θ is uniformly distributed on [0,1] and that $b(x,\theta,z) = 1$ if $x = \theta$ and 0 otherwise. Then, no matter what his strategy, the agent can never achieve adequate learning.

Theorem 3.3 Suppose that (A1), (T1) and (T2) hold, and that $\delta = 1$. Then the complete-information payoff is attained. □

It is easy to see why the complete–information optimum payoff can be approached arbitrarily closely. Indeed, our assumption that b is lower semicontinuous in x ensures that, if $\{\xi_n\}$ is dense in X, and if we pick a large but finite number N, then $M_N(\theta) = \max\{B(\xi_n,\theta) \mid 1 \leq n \leq N\}$ will be close to $M(\theta) = \sup B(\cdot,\theta)$ for most θ. Suppose therefore that the agent tries each ξ_n, $1 \leq n \leq N$, a large number of times, and then selects that ξ_n that yields the highest average payoff. Then, by the law of large numbers, he will end up with a payoff close to $M_N(\theta)$ with high probability.

To show that the complete–information payoff can actually be attained involves two further difficulties. First, it is clear that adequate information must be obtained. This requires that learning go on indefinitely. On the other hand, the agent cannot wait forever before reaping his reward. So he must make use of the partial information available to him at any given time to accumulate payoffs. Secondly, the agent must be careful that the estimate of $B(\xi_n,\theta)$ which he uses as the basis for his payoff accumulation is sufficiently accurate. Specifically, it is not enough for him simply to ensure that his estimate of $B(\xi_n,\theta)$ converges to $B(\xi_n,\theta)$ for all n. He must ensure that this convergence is uniform in n. The formal proof is given in the Appendix.

4 Existence and Robustness

In Section 3 we considered two extreme cases – that of no noise and that of no discounting. These cases are of some interest in their own right, but they are much more interesting if they are representative of the nearby cases of low noise and little discounting. This section demonstrates that the required continuity is indeed present, thereby demonstrating the relevance of the analysis of Section 3.

Continuity in the level of noise has a second advantage. With no noise, existence may fail for our model. But with a small amount of suitable noise, existence is guaranteed. So by demonstrating continuity in the level of noise, we put to rest any suspicion that the strong learning results we obtained for the case of no noise are essentially irrelevant since the premise of existence is unlikely to be fulfilled.

We begin by illustrating the non–existence problem in the case of no noise. Let $X = [-1,1]$, let $b(x,\theta)$ be defined by

$$b(x,\theta) = \begin{cases} 4-(x-1)^2 & \text{if } \theta = \theta_1 \\ 4-(x+1)^2 & \text{if } \theta = \theta_2 \end{cases}$$

and let Q assign probability 1/2 to each of θ_1 and θ_2. Then, if the agent chooses any $x \neq 0$ in period 1, he will learn the true value of the parameter. He will therefore obtain $B(x,Q) = 3 - x^2$ in period 1, and the complete–information payoff of 4 thereafter. So clearly he should pick x as close to 0 as possible. But he cannot set $x = 0$. For $x = 0$ is uninformative, and if he sets $x = 0$ in period 1 his payoff from period 2 will be at most $m(Q) = 3$. So an optimal first period strategy does not exist.

This problem can be eliminated if we introduce noise in a suitable fashion. Let $F(\cdot\,|x,\theta)$ denote the distribution of the signal y given x and θ. Then we may introduce the following assumptions:

(E1) X, Y, Z and Θ are complete separable metric spaces;

(E2) regarded as a mapping from $X \times \Theta$ to $\mathcal{PM}(Y)$ endowed with the norm topology, F is continuous;

(E3) b is bounded, Borel measurable, and upper semicontinuous in (x,θ);

(E4) X is compact;

(E1) is identical to (A1), and is purely technical. (E2) is a significantly stronger version of (A2).[13] (E3) is actually weaker than (A3). Indeed, upper semicontinuity of b in x and compactness of X (i.e. (E4)) are clearly minimal conditions for the existence of an optimal strategy.

Of the assumptions, (E3) is the fundamental assumption insofar as obtaining existence with learning is concerned. It is this assumption that ensures that there is enough noise in the model. It might arise in practice as follows: Y is $\mathbb{R}^N$; the distribution $F(\cdot\,|\,x,\theta)$ of the signal has a density $f(\cdot\,|\,x,\theta)$ with respect to Lebesgue measure on Y; $f(\cdot\,|\,x_n,\theta_n) \to f(\cdot\,|\,x,\theta)$ in mean as $(x_n,\theta_n) \to (x,\theta)$. (E2) does not involve the common assumption that, no matter what the value of (x,θ), every observation y is possible. Nor does it involve any assumption that the densities $f(y\,|\,x,\theta)$ mentioned above are jointly continuous in (x,θ,y). For example, the following case is covered: y is real valued; z is distributed uniformly on [–1,1]; there exists a continuous function $\bar{a}$ such that $a(x,\theta,z) = \bar{a}(x,\theta) + z$. This case would have been excluded if we had assumed, for example, that f was jointly continuous in all three of its arguments, or if we had made the common–support assumption.

---

[13] (A2) is equivalent to assuming that F is continuous when $\mathcal{PM}(Y)$ is endowed with the weak topology. (E2) means, more explicitly, that $\sup_A |F(A|x_n,\theta_n) - F(A|x,\theta)| \to 0$ whenever $(x_n, \theta_n) \to (x,\theta)$, where A varies over the Borel measurable sets of Y.

Theorem 4.1 Suppose that (E1)–(E4) hold. Then the agent possesses an optimal strategy.

Proof See Appendix. □

We turn now to the continuity result for noise. We define $E(v(Q_\infty))$ to be the asymptotic payoff in our model. When adequate learning occurs, the asymptotic payoff is equal to the complete–information payoff $E(M(\theta))$. So it seems reasonable to measure the departure of an optimum from adequate learning by $E(M(\theta)) - E(v(Q_\infty))$, the extent to which the asymptotic payoff falls short of the complete–information payoff. We treat only the quasiconcave case, since this illustrates the issues adequately.

Theorem 4.2 Let $X = [\underline{x},\bar{x}] \subset \mathbb{R}$. Suppose that: (i) $a \equiv b$; (ii) b is continuous in all its arguments; (iii) $\partial b/\partial x$ is continuous in all its arguments; (iv) for all θ, $b(\cdot,\theta,0)$ is quasi-concave; (v) $D(\theta) = \sup_{x,z} |\partial b/\partial x(x,\theta,z)|$ is integrable. Then the asymptotic payoff converges to the complete–information payoff as $R \to \delta_0$. □

Note that the symbol "0" simply denotes a particular element of the space Z, that δ_0 is the probability measure concentrated at 0, and that the convergence of R takes place in the weak topology. Note too that the conditions on b are stronger than those assumed for Theorem 3.2: b and $\partial b/\partial x$ are now assumed to be continuous in all their arguments, and not just in x. The proof of Theorem 4.2 is given in the Appendix.

Unlike continuity in noise, continuity in the discount factor as $\delta \to 1-$ can be obtained without assumptions additional to those made in Theorem 3.3.

Theorem 4.3 Suppose that (A1), and (T1) and (T2) of Section 3.3, hold. Then the optimal payoff converges to the complete–information optimal payoff as $\delta \to 1-$. □

The proof of Theorem 4.3 is in the Appendix. Note that it does not actually require that the optimal payoff for $\delta < 1$ be attained.

5 Inadequate Learning with a Smooth Payoff Function

In this section we present a simple example in which there is no noise and the payoff function is infinitely differentiable. In this example the agent settles with probability one on an action which he knows is not the global optimum. It highlights simultaneously the essential roles played by the assumption of analyticity in Theorem 3.1 and that of quasi-concavity in Theorem 3.2. More importantly, it shows that adequate learning is by no means an inevitable consequence of the absence of noise.

We also discuss an question raised by our example, that of which outcomes are generic. For the purposes of this discussion it will be helpful to introduce some more terminology. We have already used the phrase adequate learning to describe the case in which the agent acquires adequate knowledge asymptotically with probability one. Let us say that inadequate learning occurs when the agent acquires adequate knowledge with probability zero, and that partial learning occurs when the agent acquires adequate knowledge with probability strictly between zero and one. In this terminology, our discussion of genericity reaches the tentative conclusion that partial learning is the generic outcome. Our presentation will be informal throughout this section.

In our example the behaviour of the left–hand half of the payoff function depends only on $\varphi \in [\underline{\varphi},\overline{\varphi}]$, that of the right–hand side depends only on $\psi \in [\underline{\psi},\overline{\psi}]$, and $\theta = (\varphi,\psi)$. For any given θ, $b(\cdot,\theta)$ is fairly flat over the interval $[\underline{\varphi},\overline{\varphi}]$, attaining a local maximum of size 1 at φ. It then falls to 0 at 1.5, where all its derivatives are fixed and independent of φ. After 1.5 it remains low, except for a narrow spike in a small neighbourhood of ψ, where it attains a second local maximum of size 2. A typical such $b(\cdot,\theta)$ is pictured in Figure 1.

It should be clear that, provided δ is sufficiently small, the agent's action will converge to φ with probability one in this example. Indeed, as long as he choose an action in $[\underline{\varphi},\overline{\varphi}]$ he is guaranteed a payoff near 1. But if he chooses an action greater than 1.5 then he will almost certainly get a payoff near 0. So although choosing an action above 2 might

allow him to obtain a payoff of nearly 2 forever more, the high rate of discounting makes such a choice unattractive. Thus, with probability one, the agent's action converges to a point that he knows is not the global optimum.

Note that it is essential that the behaviour of $b(\cdot,\theta)$ over the interval [0,1.5] be determined entirely by φ, and that its behaviour over [1.5,3] be determined entirely by ψ. This ensures that the agent cannot learn anything about ψ from the outcomes of actions in $[\underline{\varphi},\overline{\varphi}]$. Yet such a situation hardly seems generic. This leads one to ask whether adequate learning is in fact the generic outcome when there is no noise.

One way of formalising genericity is as follows. Suppose that Θ and X are fixed subsets of $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively, and that we are given a density on Θ. Then we could call a property generic if it held for an open dense set of functions $b: \Theta \times X \to \mathbb{R}$. And it would appear likely that, generically, the agent achieves adequate knowledge in a finite number of periods.

This formulation may not be appropriate, however. For it implicitly restricts attention to a finite–dimensional set of possible payoff functions. (In this respect it is rather like analysing the case of a polynomial payoff function.) Yet the problem of learning about a payoff function, which is the problem in which our agent is engaged, is intrinsically infinite dimensional. It is therefore worth considering a second formulation in which a fixed set of payoff functions is given, but the agent's beliefs about these functions vary. For example, the set of payoff functions might be a subset $\Theta \subset C[0,1]$, the space of continuous functions on [0,1], and Q would then vary over $\mathcal{PM}(\Theta)$, the set of probability measures over Θ. A statement would be generic if it held for an open dense subset of $\mathcal{PM}(\Theta)$.

One problem with this second formulation is that the results may be sensitive to the choice of topology for $\mathcal{PM}(\Theta)$. Suppose first that $\mathcal{PM}(\Theta)$ is given the weak topology. Then adequate learning will be non–generic. Indeed, suppose that adequate learning occurs when the prior is Q. Let x_1 be the optimal first move, and find c such that $b(x_1,\theta) = c$

with Q–probability zero. Let $\ddot{Q}$ be the prior of an inadequate learning example in which: x_1 is a local optimum of $b(\cdot,\theta)$ with $\ddot{Q}$–probability one; $b(x_1,\theta) = c$ with $\ddot{Q}$–probability one; and optimal behaviour involves choosing x_1 forever. Then the prior $Q' = (1-\epsilon)Q + \epsilon\ddot{Q}$ leads to partial learning. For x_1 is the optimal first period action, the agent discovers which of the priors Q and $\ddot{Q}$ he really faces as soon as he observes his first-period payoff, and the former prior leads to adequate learning while the latter leads to inadequate learning. If, on the other hand, adequate learning does not occur with prior Q, then one can find a $\hat{Q}$ near Q such that $\hat{Q}$ has finite support, and such that the agent learns which of the finite number of payoff functions is actually in play as soon as he observes his first–period payoff. So inadequate learning is non–generic too.

These arguments could be taken to suggest that the question as to whether adequate learning does or does not occur simply is not the right question to ask. We would argue that this is the wrong reaction. Indeed, the problem with the weak topology is precisely that it is too weak. The sets of priors that are open in this topology are simply too big, and therefore it is relatively easy to upset any given behavioural pattern by finding another prior which is nearby according to the weak topology, but which is in fact quite dissimilar. Suppose therefore that $\mathcal{PM}(\Theta)$ is endowed with the norm topology. Then the same argument as in the previous paragraph shows that adequate learning is non–generic. But the argument showing that inadequate learning is non–generic breaks down. Indeed, it seems very likely that all priors in a neighbourhood of a prior that leads to partial learning themselves lead to partial learning. And if this is so then partial learning will be generic.[14]

---

[14] There is a significant problem, however. When 'generic' is given the meaning we are using here, non–existence is probably the generic outcome. To see this, simply note that any prior for which existence does obtain can be mixed with a small amount of a prior for which it does not in such a way as to destroy the existence. Moreover, non–existence probably obtains for all priors in a neighbourhood of the resulting 'mixed' prior. So in order to make our claim that partial learning is generic precise, one would either have to give a meaning to terms such as adequate and inadequate learning in the case where non–existence obtains, or consider genericity in the topology induced by the norm topology on the set of priors for which existence does obtain.

To summarise, we think that the question of genericity is an interesting one in our model. Great care must, however, be taken to arrive at a sensible definition of genericity. Of the definitions we have discussed, we think that the last, which uses the norm topology on the space of priors, appears to be the most appropriate. If this is correct, then it would appear that partial learning is the only generic outcome.

6 Inadequate Learning with a Discontinuous Payoff Function

In this section we provide a second example in which there is inadequate learning. The example has a number of attractive features. First, an optimum exists in spite of the absence of noise. Secondly, inadequate learning occurs with probability one. Thirdly, there is a positive probability that learning ceases altogether after a finite number of periods, and a positive probability that learning goes on forever.

Each period a monopolist tries to sell an indivisible good to a consumer. The monopolist sets a price x, and the consumer buys iff the price is less than or equal to his reservation price θ. Initially the monopolist believes that θ is distributed uniformly on [0,1], but he revises these beliefs in the light of his failure or success in selling at the price he asks.

For this problem: $\Theta = [0,1]$; $X = [0,1]$; $b(x,\theta) = x$ if $x \leq \theta$ and $b(x,\theta) = 0$ if $x > \theta$; and $a \equiv b$. So the standing assumptions (A2) and (A3) are not satisfied. As a result, we must pay attention to the specific structure of our problem.

Because the monopolist can only observe whether he sells or not, his beliefs in period $t + 1$ must be that θ is distributed uniformly in an interval $[\underline{\theta}_t, \overline{\theta}_t]$. Here $\underline{\theta}_t$ is the highest price at which he has so far sold, or 0 if he has not yet succeeded in selling; and $\overline{\theta}_t$ is the lowest price at which he has failed to sell, or 1 if he has not yet failed to sell. Clearly $\underline{\theta}_t$ is non–decreasing in t and $\overline{\theta}_t$ is non–increasing. Moreover we need not calculate the value function v for all possible priors. Indeed, we need only calculate v in the case where his beliefs are that θ is uniformly distributed on [g,h]. To this end, consider the Bellman equation

$$w(g,h) = \sup_{0 \leq \lambda \leq 1} \Big[(1-\delta)(1-\lambda)(g + \lambda(h-g))$$

$$+ \delta(1-\lambda)w(g+\lambda(h-g),h) + \delta\lambda w(g,g+\lambda(h-g))\Big] \qquad \ldots(6.1),$$

which is a simplified version of the original equation (2.3).

Lemma 6.1 Equation (6.1) has a unique solution. This solution is continuous and convex.

Proof It is easy to check that continuity is preserved by the Bellman operator. Similarly, since the supremum of a family of convex functions is convex, convexity is preserved too. The lemma therefore follows from standard considerations. □

Theorem 6.1 There exists an optimal strategy.

Proof Since w is continuous, the supremum on the right–hand side of (6.1) is attained. Indeed, the correspondence mapping (g,h) into those λ which maximise the right–hand side of (6.1) is upper semicontinuous. It follows that a measurable selection from this correspondence exists. Standard considerations then show that this selection generates an optimal strategy, and that w is the value function for the problem under consideration. □

Since w(g,h) is the value function for the problem, it must be homogeneous of degree 1 (because the problem is invariant under scaling), and it must be strictly increasing in both arguments. The first of these properties is especially useful, since it shows that we can confine our analysis to the case h = 1. Let $W(g) = w(g,1)$. In order to proceed further, it is helpful to know the circumstances in which learning can cease. Now no further information is obtained if a price x_{t+1} such that $x_{t+1} \leq \underline{\theta}_t$ or $x_{t+1} \geq \overline{\theta}_t$ is chosen. And of these possibilities only $x_{t+1} = \underline{\theta}_t$ could possibly be optimal. So learning can cease iff $\underline{\theta}_t$ is an optimal price.

Lemma 6.2 Let $c = 1/(2 - \delta)$. Then learning can cease iff $\underline{\theta}_t/\overline{\theta}_t \geq c$. In this case $\underline{\theta}_t$ is the unique optimal price.

Proof Suppose that g is an optimal price when beliefs are summarised by (g,1). Then setting a price of g forever is an optimal strategy, and so $W(g) = g$. Conversely, if $W(g) = g$ then g is an optimal price when beliefs are given by (g,1). So, overall, learning can cease iff $W(g) = g$.

Now setting g forever is always a possible strategy. So $W(g) \geq g$. Also $W(1) = 1$. Since W is convex, it follows that the set of g such that $W(g) = g$ is a closed interval with right-hand endpoint 1. Hence $W(g) = g$ implies that g solves the problem

$$\max_{g \leq x \leq 1} (1 - \delta)\left[\frac{1-x}{1-g}\right]x + \delta\left[\frac{1-x}{1-g}\right]\cdot + \delta\left[\frac{x-g}{1-g}\right]g \qquad \ldots(6.2)$$

which in turn implies that $g \geq c$. Conversely, if $\tilde{g} \geq c$ then g solves (6.2) for all $g \geq \tilde{g}$, which implies that $W(\tilde{g}) = \tilde{g}$. So $W(g) = g$ iff $g \geq c$. Finally, the solution to (6.2) is always unique. □

Lemma 6.2 tells us that if $\underline{\theta}_t \geq c\overline{\theta}_t$ then next period's price $x_{t+1} = \underline{\theta}_t$, and that if $\underline{\theta}_t < c\overline{\theta}_t$ then $x_{t+1} > \underline{\theta}_t$. Our next goal is to show that, in this latter case, $x_{t+1} < c\overline{\theta}_t$ as well. This turns out to be a surprisingly difficult result. The first step is to show that $x_{t+1} \leq c\overline{\theta}_t$.

Lemma 6.3 Suppose that $\underline{\theta}_t < c\overline{\theta}_t$. Then $x_{t+1} \leq c\overline{\theta}_t$.

The intuition behind this result is as follows. Consider an $x > \overline{\theta}_t c$. If the agent were informed whether $\theta < \overline{\theta}_t c$ or $\theta \geq \overline{\theta}_t c$, he would weakly prefer $c\overline{\theta}_t$ to x in the first case, and strictly prefer $c\overline{\theta}_t$ to x in the second. Hence x is dominated by $\overline{\theta}_t c$. This intuition must, however, be treated with care.

Proof It suffices to consider the case where $(\underline{\theta}_t, \overline{\theta}_t) = (g,1)$, and to show that, in this case, the optimal x is at most c. To this end, note that the payoff from playing $x > c$ now and optimally thereafter is at most equal to the payoff of playing x now, then being informed whether $\theta \in [g,c)$ or $\theta \in [c,1]$ (in addition to the usual information as to whether $\theta \in [x,1]$ or $\theta \in [g,x)$), and then playing optimally thereafter. The latter payoff is

$$\frac{1-c}{1-g}\left[(1-\delta)\left[\frac{1-x}{1-c}\right]x + \delta\left[\frac{1-x}{1-c}\right]w(x,1) + \delta\left[\frac{x-g}{1-c}\right]w(c,x)\right] + \frac{c-g}{1-g}w(g,c)$$

$$< \frac{1-c}{1-g}\left[(1-\delta)cd + \delta w(c,1)\right] + \frac{c-g}{1-g}w(g,c)$$

(since c is strictly optimal when it is known that $\theta \in [c,1]$)

$$= (1-\delta)\left[\frac{1-c}{1-g}\right]c + \delta\left[\frac{1-c}{1-g}\right]w(c,1) + \delta\left[\frac{c-g}{1-g}\right]w(g,c),$$

(which is the payoff from playing c now and optimally thereafter). That is, overall c is a strictly better choice than any $x > c$. □

It will be seen from the proof just how slippery was the intuition with which we started. We were only able to make it precise by feeding the agent the extra information after he made his decision, and by noting that the extra information turned out to be redundant.

Obtaining the strict inequality $x_{t+1} < c\overline{\theta}_t$ is significantly harder. The basic difficulty can, however, be understood in terms of our intuition. To prove that $x_{t+1} < c\overline{\theta}_t$ we have to show that there is some $x < c\overline{\theta}_t$ that dominates the choice of $c\overline{\theta}_t$ itself. It is easy to see that some $x < c\overline{\theta}_t$ is better than $c\overline{\theta}_t$ when $\theta < c\overline{\theta}_t$. The problem is that any $x < c\overline{\theta}_t$ is strictly worse than $c\overline{\theta}_t$ when $\theta \geq c\overline{\theta}_t$. So we have to show that the trade-off between optimality in the two events $\theta < c\overline{\theta}_t$ and $\theta \geq c\overline{\theta}_t$ favours a compromise between the two.

Lemma 6.4 Suppose that $\underline{\theta}_t < c\overline{\theta}_t$. Then $x_{t+1} < c\overline{\theta}_t$.

Proof In the first step of the proof we show that the right-hand derivative of $(h-g)w(g,h)$ with respect to h is at most h. In the second, we exploit the envelope theorem to show that this upper bound can be refined progressively, obtaining a final bound of $\max\{g,ch\}$. The last step of the proof uses this bound to show that, when $g < ch$, the choice of action ch is dominated by some $x < ch$.

Turning to the first step, note that $\partial^+((h-g)w(g,h))/\partial h$ is well defined because w is convex in h. Next,

$$(h+\epsilon-g)w(g,h+\epsilon-g) \leq (h-g)w(g,h) + \epsilon w(h,h+\epsilon),$$

by convexity of the value function w. Moreover $w(h,h+\epsilon) \leq h+\epsilon$. So

$$\frac{(h+\epsilon-g)w(g,h+\epsilon-g)-(h-g)w(g,h)}{\epsilon} \leq h+\epsilon.$$

The required inequality

$$\frac{\partial^+}{\partial h}((h-g)w(g,h)) \leq h \qquad \ldots(6.3)$$

follows on letting $\epsilon \to 0+$.

For the second step, we must distinguish between the cases $g = h$ and $g < h$. If $g = h$ then $\partial^+((h-g)w(g,h))/\partial h \leq h = g$ by the first step, and the required inequality follows trivially. If $g < h$, let x be any optimal choice of action when $\theta \in [g,h]$. Then $x < h$ and we may apply the envelope theorem to the Bellman equation to conclude that

$$\frac{\partial^+}{\partial h}((h-g)w(g,h)) \leq (1-\delta)x + \delta\frac{\partial^+}{\partial h}((h-x)w(x,h)) \qquad \ldots(6.4).$$

But $x \leq \max\{g,ch\}$ by Lemmas 6.2 and 6.3, and $\partial^+((h-x)w(x,h))/\partial h \leq h$ by (6·3). So (6.4) implies that

$$\frac{\partial^+}{\partial h}((h-g)w(g,h)) \leq (1-\delta)\max\{g,ch\} + \delta h \qquad \ldots(6.5),$$

which is an improved version of (6.3). Since this inequality holds for all (g,h) such that $g < h$, we may apply it to the right-hand side of (6.4) to conclude that

$$\frac{\partial^+}{\partial h}((h-g)w(g,h)) \leq (1-\delta)x + \delta((1-\delta)\max\{x,ch\} + \delta h)$$

$$\leq (1-\delta)(1+\delta)\max\{x,ch\} + \delta^2 h$$

$$\leq (1-\delta)(1+\delta)\max\{g,ch\} + \delta^2 h.$$

Iterating this argument yields the desired conclusion.

For the third step, suppose that $g < ch$, and consider $f(x) = (1-\delta)(h-x)x + \delta(h-x)w(x,h) + \delta(x-g)w(g,x)$ ($f/(h-g)$ is the maximand in the Bellman equation). To show that the optimal x is less than ch, it will suffice to show that $\partial^- f(x)/\partial x|_{x=ch} < 0$. But

$$\frac{\partial^- f(x)}{\partial x}\Big|_{x=ch} \leq \frac{\partial^+ f(x)}{\partial x}\Big|_{x=ch}$$

by convexity of w. Moreover $f(x) = (h-x)x + \delta(x-g)w(g,x)$ if $ch \leq x < h$. So

$$\frac{\partial^+ f(x)}{\partial x} \leq h - 2x + \delta \max\{g,cx\} \qquad \text{(for such x)}$$

$$= -\delta(ch - \max\{g,c^2h\}) \qquad \text{(when } x = ch\text{)}$$

$$< 0.$$

This completes the proof of the Lemma. □

We are now in a position to prove the promised results about learning behaviour. Note that the price falls in period t (i.e. $x_t < x_{t-1}$) iff the monopolist fails to sell in period $t-1$.

Theorem 6.2 Suppose that $\theta = 0$. Then the price falls in every period, and converges to zero as $t \to \infty$. In particular, the monopolist learns the true value of the parameter.

Proof Certainly $\underline{\theta}_0 = 0$ and $\overline{\theta}_0 > \underline{\theta}_0$. If $\underline{\theta}_{t-1} = 0$ and $\overline{\theta}_{t-1} > 0$ then the monopolist will choose $0 < x_t < c\overline{\theta}_{t-1}$ by Lemma 6.3. Since $\theta = 0$ he will fail to sell at x_t. So $\underline{\theta}_t = 0$ and $\overline{\theta}_t = x_t < c\overline{\theta}_{t-1}$. Hence both the price and the upper bound on the consumer's reservation price converge to zero geometrically. □

Theorem 6.3 Suppose that $\theta > 0$. Then learning is inadequate.

Proof Let σ be the first period in which the monopolist makes a sale, and let $\rho_t = \underline{\theta}_t/\overline{\theta}_t$. Since $\theta > 0$, $\sigma < \infty$.

Now $\rho_t = 0$ for all $t < \sigma$. Also, $\underline{\theta}_\sigma = x_\sigma$ and $\overline{\theta}_\sigma = x_{\sigma-1}$. Hence $\underline{\theta}_\sigma < c\overline{\theta}_\sigma$, and $0 < \rho_\sigma < c$. Thirdly, for any t, if $0 < \rho_t < c$ then $\rho_t < \rho_{t+1} < 1$. (In particular, $\rho_\sigma < \rho_{\sigma+1} < 1$.) Fourthly, if $\rho_t \geq c$ then $\rho_{t+1} = \rho_t$.

Overall, then, $\{\rho_t\}$ is a non–decreasing sequence that converges to a limit $\rho_\infty < 1$. Since $\underline{\theta}_t/\overline{\theta}_t \leq \rho_\infty$ for all t, adequate learning cannot occur. □

Note that the proof of Theorem 6.3 also tells us that there is at least one period in which the monopolist raises his price. That is, there exists $t > 1$ such that $x_t > x_{t-1}$.[15]

Our next result makes more precise the nature of the inadequate learning signaled by Theorem 6.3. It shows that learning may go on forever, or that it may cease altogether after a finite number of periods.

Theorem 6.3 Suppose that $\theta > 0$. Then the number of periods in which the price falls is finite. (It may be zero.) If τ is the last period in which the price falls then the price either remains constant at x_τ forever more (i.e. $x_t = x_\tau$ for all $t \geq \tau$), or it is strictly increasing

---

15 This contrasts with the results of Lazear (1986), who finds that prices must fall. But that is only to be expected in a model without repeat purchases.

forever more (ie, $x_{t+1} > x_t$ for all $t \geq \tau$). Moreover each of these possibilities occurs with positive probability.

Note that τ is a random variable. In the case when price remains constant, the monopolist knows that τ has been reached. But in the case when price increases forever, he does not. (For all he knows, his current price increase could result in a failure to sell.) So τ is not a Markov time.

Proof Suppose that the price falls in period t. Then $\overline{\theta}_{t-1} = x_{t-1} \leq c\overline{\theta}_{t-2}$. So if the price falls in an infinite number of periods then $\overline{\theta}_t \to 0$ as $t \to \infty$. This contradicts the assumption that $\theta > 0$. Hence $\tau < \infty$. There are now two possibilities. If $\rho_{\tau-1} \geq c$ then $x_t = \underline{\theta}_{\tau-1}$ for all $t \geq \tau$ by Lemma 6.2. If, on the other hand, $\rho_{\tau-1} < c$ then $\underline{\theta}_{\tau-1} < x_\tau < \overline{\theta}_{\tau-1}$ by Lemma 6.3. Also, by definition of τ, the monopolist will succeed in selling in period τ, so $\underline{\theta}_\tau = x_\tau > \underline{\theta}_{\tau-1}$ and $\rho_\tau = \underline{\theta}_\tau/\overline{\theta}_\tau = x_\tau/\overline{\theta}_{\tau-1} < c$. Moreover this argument can be continued indefinitely. So the price increases strictly forever. (Note how we have exploited the foreknowledge obtained from the non–Markov time τ.)

It remains to show that both of these possibilities can arise with positive probability. It is easy to see that this is true of the second possibility: it occurs whenever $\theta \geq c$. To obtain the second, let $\{x_t \mid t_t \geq 1\}$ be the sequence of prices charged when $\theta \geq c$. Then $x_t \to c-$. (If it did not, then it would follow from the upper semicontinuity of the set of maximisers of the right–hand side of (6.1) that x_∞ is an optimal price when $(g,h) = (x_\infty,1)$, where $x_\infty < c$ is the limit of the x_t. But this contradicts Lemma 6.2.) So there exists T such that $x_T \geq c^2$. The first possibility arises whenever $\theta \in [x_T,c)$. □

It should be emphasized that, even when learning goes on forever, it does not result in adequate knowledge.

7. Conclusion

In this last section we briefly summarize and interpret the main findings of the paper.

Our first set of results concerns the long–run learning behaviour of an economic agent facing an unknown payoff function. Our main convergence result, which states that in the long run the benefit of experimentation tends to zero, provides a general perspective applicable both to the case where the payoff function is deterministic and to the case where it is stochastic. In both cases the motive of learning induces smaller and smaller deviations from the myopic optimum as time goes to infinity. In other words, whatever is learned asymptotically from experimentation can only be learned as a result of local experimentation about the myopic optimum. In particular, the possibility or otherwise of adequate learning relates directly to local properties of the payoff function.

Our second set of results relates to the case where the payoff function is deterministic. Here we have shown that, when the payoff function is, in addition, analytic, local experimentation allows for global extrapolation, and that the agent is therefore able to attain the true global optimum asymptotically. On the other hand, if the payoff function is infinitely differentiable but not analytic, an example shows that this result may break down. Similarly, when the payoff function is continuously differentiable, local experimentation allows for local extrapolation, by providing an arbitrarily precise estimate of the slope of the true payoff function at any given point. This in turn guarantees that local experimentation will eventually lead to a local optimum of the payoff function. On the other hand, if the payoff function is continuous but not continuously differentiable, an example (which we do not provide in the paper) shows that this result may break down. (A case of special interest here is that in which the payoff function is continuously differentiable and quasiconcave. For in that case any local optimum is also a global optimum, and the global optimum will be reached as a result of local extrapolation.)

This second set of results is of interest for at least four reasons. First, it represents a partial (though fairly extensive) characterisation of those situations in which adequate learning will and will not occur. This is in contrast with much of the literature, which has tended to concentrate on providing examples of inadequate learning. Secondly, this characterisation illustrates the principle which emerged from the first set of results, namely that local properties of the payoff function will be crucial in determining whether adequate learning occurs or not. Indeed, the central properties of the payoff function are analyticity, differentiability and continuity, all of which are local properties. Thirdly, as we illustrated in our discussion of robustness, this set of results extends (at least approximately) to the case of a small amount of noise. Fourthly, the case of a deterministic payoff function occurs frequently in applications.

By analysing a polar case of special interest, the second set of results illustrates some of the factors that will determine whether adequate learning will or will not occur in the general case of a stochastic payoff function. But other factors can be expected to come into play in this case. For example, although local experimentation is likely to continue to be of central importance, the nature of learning from local experimentation is likely to change. In particular, if small local changes in action are to provide usable information about the shape of the payoff functions, they may have to be maintained over many periods. This and similar effects could render learning more difficult, and make adequate learning less likely. On the other hand, as we know from one of the convergence results, the agent will ultimately learn the true distribution of his observations. This could make adequate learning easier. Indeed, suppose that the agent's payoff is a noisy function of his action and the true parameter, and that he simply observes his payoff in each period. Then, in the limit, he will know the true payoff distribution associated with his limiting action. This is at least as much information as he would obtain in the corresponding deterministic case — in which he would learn only the expectation of the true payoff distribution — and is likely to be significantly more.

Our third and last set of results relates to the problem of characterising the process of adjustment to the long–run outcome, as opposed to the problem of characterising the long–run outcome itself. This problem is much harder to deal with in the general case, so we worked instead with a simple example in which a monopolist made repeat sales to a myopic buyer. In this example we were able to build up a detailed qualitative picture of the optimal strategy, and of the associated process of adjustment. It seems likely that further results on short–run learning behaviour will be obtained similarly in the context of simple examples. One possibility here would be to extend our example of a monopolist to the case where the buyer behaves strategically. Another would be to analyse the interaction between competition and learning in an oligopolistic model.

Appendix

This appendix begins by providing the justification for the material in Section 2. In doing so, we follow a pattern of development slightly different from that followed in Section 2. It continues with a proof of the existence theorem of Section 4; and concludes by collecting together all the remaining proofs missing from Sections 3 and 4.

For our development of the material of Section 2 we need the following two assumptions:

(A2′) a is Borel measurable;

(A3′) b is bounded and Borel measurable.

These assumptions differ from (A2) and (A3) in that they do not assume that a and b are continuous in (x,θ). We assume henceforth that (A1), (A2′) and (A3′) hold. We also assume that all spaces of probability measures are endowed with the weak topology unless explicitly stated to the contrary.

Suppose now that an arbitrary strategy s is given. Let $\mathcal{F}_0$ be the trivial σ–algebra. Next, for all $t \geq 1$ the agent will have observed $(y_1,\ldots,y_t)$ at the beginning of stage $t+1$. In this case let $\mathcal{F}_t$ be the σ–algebra generated by $(y_1,\ldots,y_t)$. Finally, let $\mathcal{F}_\infty$ be the σ–algebra generated by the sequence $(y_1,y_2,\ldots)$. Then standard considerations show that, for each $0 \leq t \leq \infty$, we may construct an r.c.p.d. (regular conditional probability distribution) of P given $\mathcal{F}_t$. That is, there exists a P_t satisfying:

(i) for all $\omega \in \Omega$, $P_t(\cdot\,|\,\omega) \in \mathcal{PM}(\Omega)$;

(ii) $P_t(A\,|\,\cdot)$ is $\mathcal{F}_t$–measurable for all $A \in \mathcal{F}$;

(iii) $P(A \cap B) = \int_B P_t(A \mid \omega) dP(\omega)$ for all $A \in \mathcal{F}$ and all $B \in \mathcal{F}_t$;

(iv) $P_t(\cdot \mid \omega)$ is concentrated on the atom[16] of $\mathcal{F}_t$ containing ω for P–almost all ω

$P_t(\cdot \mid \omega)$ is to be interpreted as the agent's posterior about the state of the world, based on the information available to him prior to stage $t + 1$, when the true state of the world is ω. (i) and (iii) capture the idea that $P_t(\cdot \mid \omega)$ is the agent's posterior. (ii) captures the idea that this posterior depends only on information available to him prior to stage $t + 1$. Finally, (iv) ensures that his posterior assigns probability zero to the set of states of the world that do not generate his observations to date.

The agent's posterior concerning the parameter of interest are now given by $Q_t(\cdot \mid \omega)$, the marginal of $P_t(\cdot \mid \omega)$ over Θ. We regard Q_t as a random variable defined on Ω and taking values in $\mathcal{PM}(\Theta)$. With this convention, the following result is standard.

Theorem A.1 $\{Q_t \mid 1 \leq t \leq \infty\}$ is a martingale. Moreover $Q_t \to Q_\infty$ a.s. as $t \to \infty$–. □

Note that $\{Q_t \mid 1 \leq t \leq \infty\}$ follows a martingale in the sense that there exists a single null set outside which $E[\int f dQ_t \mid \mathcal{F}_u] = \int f dQ_u$ for all $1 \leq u \leq t \leq \infty$ and all bounded continuous $f: \Theta \to \mathbb{R}$. Also, by definition of convergence in the weak topology, there exist a single null set outside which $\int f dQ_t \to \int f dQ_\infty$ for all such f.

At this point we remind the reader that, by definition,

$$m(Q) = \sup_s E[\pi_1]$$

and

---

16 The atom of $\mathcal{F}_t$ containing ω is the smallest $\mathcal{F}_t$–measurable set A such that $\omega \in A$.

$$v(Q) = \sup_s E\left[(1-\delta)\Sigma_{t=1}^{\infty}\delta^{t-1}\pi_t\right].$$

We also introduce the following assumption:

(B) there exists a countable set of strategies S such that, for all $Q \in \mathcal{PM}(\Theta)$, $\sup_{s\in S} E[(1-\delta)\Sigma_{t=1}^{\infty}\delta^{t-1}\pi_t] = v(Q)$.

We can now state two results, both of which follow from standard considerations (cf. Striebel [1975], for example).

Theorem A.2 Suppose that (B) holds. Then $\{v(Q)\,|\,1 \le t \le \infty\}$ follows a submartingale. □

The essential point here is that v can be expressed as the upper envelope of a countable number of linear functions by (B), and that Jensen's inequality therefore applies.

Theorem A.3 Suppose that (B) holds, and let s be optimal. Then $v(Q_t) = E[(1-\delta)\,\Sigma_{u=1}^{\infty}\delta^{u-1}\pi_{t+u}\,|\,\mathcal{F}_t]$ a.s. for all $t \ge 1$. □

The point here is that, while it is obvious that $v(Q_t) \ge E[(1-\delta)\,\Sigma_{u=1}^{\infty}\,\delta^{u-1}\pi_{t+u}\,|\,\mathcal{F}_t]$, assumption (B) is needed to show that we can construct a strategy that improves on s if this inequality holds strictly on a set of positive probability.

Theorem A.4 Suppose that (B) holds, and that s is optimal. Then $v(Q_t) - m(Q_t) \to 0$ a.s. as $t \to \infty$.

Proof We have

$$v(Q_t) = E[(1-\delta)\sum_{u=1}^{\infty} \delta^{u-1}\pi_{t+u} | \mathcal{F}_t] \qquad \text{a.s.}$$

(by Theorem A.3)

$$= (1-\delta)E[\pi_{t+1} | \mathcal{F}_t] + \delta E[(1-\delta)\sum_{u=1}^{\infty} \delta^{u-1}\pi_{t+1+u} | \mathcal{F}_{t+1} | \mathcal{F}_t] \text{ a.s.}$$

$$= (1-\delta)\int \pi_{t+1} dP_t + \delta E[v(Q_{t+1}) | \mathcal{F}_t] \qquad \text{a.s.}$$

(because $\int \pi_{t+1} dP_t$ is a version of the conditional expectation $E[\pi_{t+1} | \mathcal{F}_t]$, and by Theorem A.3 again). Moreover

$$\int \pi_{t+1} dP_t = \int b(x_{t+1}(\tilde{\omega}), \tilde{\theta}, \tilde{z}_{t+1}) dP_t(\tilde{\omega} | \omega)$$

(by definition)

$$= \int b(x_{t+1}(\omega), \tilde{\theta}, \tilde{z}_{t+1}) dP_t(\tilde{\omega} | \omega)$$

(by property (iv) of an r.c.p.d., and the fact that x_{t+1} is $\mathcal{F}_t$–measurable)

$$= \int B(x_{t+1}(\omega), \tilde{\theta}) dQ_t(\tilde{\theta} | \omega)$$

(because z_{t+1} is independent of θ conditional on $\mathcal{F}_t$).

$$= B(x_{t+1}(\omega), Q_t(\cdot | \omega))$$

(by definition of B)

$$\leq m(Q_t(\cdot \mid \omega))$$

(by definition of m). Hence

$$v(Q_t) - m(Q_t) \leq \frac{\delta}{1-\delta}\left[E[v(Q_{t+1}) \mid \mathcal{F}_t] - v(Q_t)\right] \text{ a.s.}$$

But $\{v(Q_t) \mid 1 \leq t \leq \infty\}$ is a bounded submartingale. Standard martingale results therefore show that there exists a bounded $\bar{V}$ such that $v(Q_t) \to \bar{V}$ a.s., $v(Q_t) \leq E[\bar{V} \mid \mathcal{F}_t]$ a.s. for all t, and $E[\bar{V} \mid \mathcal{F}_t] \to \bar{V}$ a.s. It follows at once that

$$E[v(Q_{t+1}) \mid \mathcal{F}_t] - v(Q_t) \leq E[\bar{V} \mid \mathcal{F}_{t+1} \mid \mathcal{F}_t] - v(Q_t) \text{ a.s.}$$

$$= E[\bar{V} \mid \mathcal{F}_t] - v(Q_t) \qquad \text{a.s.}$$

$$\to 0 \qquad \text{a.s.}$$

This completes the proof. □

So far we have phrased our analysis in terms of the non–basic assumption (B). This potential weakness will be rectified shortly: we shall show that, to obtain (B), it is sufficient to assume that a is continuous in (x,θ) and that b is lower semicontinuous in (x,θ). The reason why we have not invoked these assumptions before is that we feel that Theorems A.1 to A.4 have a greater generality than such continuity assumptions on a and b might suggest.

Theorem A.5 Suppose that a is continuous in (x,θ) and that b is lower semicontinuous in (x,θ). Then v and m are both lower semicontinuous in Q.

The claim that m is lower semicontinuous is easily verified, and depends in fact only on the assumption that b is lower semicontinuous. The claim that v is lower semicontinuous is much harder to prove. Indeed, in order to prove it we need to extend the concept of a strategy for period t + 1 from a Borel measurable function $s_{t+1}: Y^t \to X$ to that of a Borel measurable function $s_{t+1}: Y^t \to \mathcal{PM}(X)$. The bulk of the proof is accounted for by the following three lemmas.

Lemma A.1 Suppose that $s_{t+1}: Y^t \to \mathcal{PM}(X)$ is Borel measurable, and let a $\lambda^t \in \mathcal{PM}(Y^t)$ be given. Then, for all $\epsilon > 0$, there exists a continuous $s^{\epsilon}_{t+1}: Y^t \to \mathcal{PM}(X)$ such that $d(s^{\epsilon}_{t+1}(y), s_{t+1}(y)) < \epsilon$ for all y in a set of λ^t–measure $1 - \epsilon$.

In the statement of the lemma, $d(\cdot,\cdot)$ is any metric on $\mathcal{PM}(X)$ which induces the weak topology.

Proof In order to simplify notation, suppose simply that $s: Y \to \mathcal{PM}(X)$ is Borel measurable, $\lambda \in \mathcal{PM}(Y)$, and $\epsilon > 0$ is given.

Let $\{p_n \mid n \geq 1\}$ be a sequence that is dense in $\mathcal{PM}(X)$. Let $B_{\epsilon}(p_n)$ denote the ball of radius ϵ centred on p_n. Define $\bar{s}: Y \to \mathcal{PM}(X)$ inductively by the formula: $\bar{s}(y) = p_1$ if $s(y) \in B_{\epsilon}(p_1)$; $\bar{s}(y) = p_{n+1}$ if $s(y) \in B_{\epsilon}(p_{n+1}) \backslash \cup_{i=1}^{n} B_{\epsilon}(p_i)$. Because $\{p_n\}$ is dense, $\{B_{\epsilon}(p_n)\}$ covers $\mathcal{PM}(X)$, and $\bar{s}$ is well defined. Moreover it is clear that $d(\bar{s}(y), s(y)) < 2\epsilon$ for all y.

Next, let $E_n = \bar{s}^{-1}(p_n)$ for all n. Find N such that $\Sigma_{n=1}^{N} \lambda(E_n) > 1 - \epsilon$. For each $n \leq N$, find $K_n \subset E_n$ such that K_n is compact and $\lambda(E_n \backslash K_n) < \epsilon/N$. (Such K_n exist because every Borel measure on a complete separable metric space is tight. See

Parthasarthy (1967; Chapter 3) for example.) Find disjoint open sets V_n such that $K_n \subset V_n$ for all $n \leq N$. Find continuous functions $\chi_n: Y \to [0,1]$ such that $\chi_n = 1$ on K_n and $\chi_n = 0$ outside V_n. (Such functions exist by Urysohn's Lemma.) Let $\psi_1 = \chi_1$; for each $2 \leq n \leq N$ let $\psi_n = (1-\chi_1)\ldots(1-\chi_{n-1})\chi_n$; and let $\psi_{N+1} = 1 - \Sigma_{n=1}^{N} \psi_n$. And define $s^\epsilon = \Sigma_{n=1}^{N+1} p_n \psi_n$.

It can be checked that $s^\epsilon(y) = p_n$ on K_n, and therefore that $d(s^\epsilon(y), s(y)) < 2\epsilon$ on a set of λ-measure at least $1 - 2\epsilon$. Also s^ϵ is continuous by construction. Finally, $\psi_n \geq 0$ for all $1 \leq n \leq N+1$, and $\Sigma_{n=1}^{N+1} \psi_n = 1$. So $s^\epsilon(y) \in \text{conv}\{p_1, p_2, \ldots, p_{N+1}\} \subset \mathcal{PM}(X)$ for all y. □

Let s be any fixed strategy, and let $v_s(Q)$ be the payoff obtained when s is played and priors are Q.

Lemma A.2 Suppose that a is continuous in (x,θ) and that b is lower semicontinuous in (x,θ). Let s be any continuous strategy. Then v_s is lower semicontinuous.

By saying that s is continuous, we mean that $s_{t+1}: Y^t \to \mathcal{PM}(X)$ is continuous for all $t \geq 1$. (s_1 is just an element of $\mathcal{PM}(X)$.)

Proof For each $(x,\theta) \in X \times \Theta$, let $\alpha(\cdot\,|\,x,\theta)$ be the distribution of the random variable $a(x,\theta,\cdot): Z \to Y$ obtained when Z is given distribution R. Because a is continuous in (x,θ), $\alpha(\cdot\,|\,\cdot)$ is a continuous transition probability from $X \times \Theta$ to Y. Let $\sigma_t(\cdot\,|\,y^t) = s_{t+1}(y^t)$ for all $y^t \in Y^t$. If s is continuous, σ_t is likewise a continuous transition probability from Y^t to X for all $t \geq 1$.

Now suppose we are given priors Q and a strategy s. Then we may build a sequence of probability distributions $\mu^t(\cdot\,|\,s,Q)$ over $\Theta \times X^t \times Y^t$ inductively as follows. First combine the marginal $Q \otimes s_1$ over $\Theta \times X$ with the transition probability which takes (θ,x)

into $\delta_{(\theta,x)} \otimes \alpha(\cdot\,|x,\theta)$ to obtain $\mu^1(\cdot\,|s,Q)$ over $\Theta \times X \times Y$. Then, having obtained the distribution $\mu^t(\cdot\,|s,Q)$ over $\Theta \times X^t \times Y^t$ for some $t \geq 1$, apply the transition probability that takes (θ,x^t,y^t) into $\delta_\theta \otimes \delta_{x^t} \otimes \sigma_t(\cdot\,|y^t) \otimes \delta_{y^t}$ followed by the transition probability which takes (θ,x^{t+1},y^t) into $\delta_\theta \otimes \delta_{x^{t+1}} \otimes \delta_{y^t} \otimes \alpha(\cdot\,|x_{t+1},\theta)$, and obtain the distribution $\mu^{t+1}(\cdot\,|s,Q)$ over $\Theta \times X^{t+1} \times Y^{t+1}$. The sequence of distributions obtained in this way is consistent, so there is a unique $\mu^\infty(\cdot\,|s,Q)$ over $\Theta \times X^\infty \times Y^\infty$ of which they are the marginals. Moreover, if s is continuous then each $\mu^t(\cdot\,|s,Q)$ depends continuously on Q, and therefore $\mu^\infty(\cdot\,|s,Q)$ too depends continuously on Q.

Next, let $B(x,\theta) = \int b(x,\theta,z)dR(z)$ for all (x,θ). B is bounded and lower semicontinuous in (x,θ) because b is. Let $\beta(x^\infty,\theta) = (1-\delta)\,\Sigma_{t=1}^{\infty}\,\delta^{t-1}B(x_t^\infty,\theta)$ for all $(x^\infty,\theta) \in X^\infty \times \Theta$. Because B is bounded and lower semicontinuous, so is β. Finally, note that

$$v_s(Q) = \int \beta(x^\infty,\theta)\, d\mu^\infty(\theta,x^\infty,y^\infty\,|s,Q)$$

for all Q. Because $\mu^\infty(\cdot\,|s,Q)$ is continuous in Q, and because β is bounded and lower semicontinuous, v_s is bounded and lower semicontinuous. □

Lemma A.3 Suppose that a is continuous in (x,θ) and that b is lower semicontinuous in (x,θ). Let s be any strategy. Then $v_s(Q) \leq \sup_{\bar{s}} v_{\bar{s}}(Q)$, where $\bar{s}$ varies over all continuous strategies.

Proof For the proof we shall need some notation. First, if s is a strategy then let $s^t = (s_1,s_2,\ldots,s_t)$ and ${}^{t+1}s = (s_{t+1},s_{t+2},\ldots)$. Secondly, suppose that we are given $(Q,x^t,y^t) \in \mathcal{PM}(\Theta) \times X^t \times Y^t$, $\xi \in \mathcal{PM}(X)$, and a strategy s. Then we may construct a probability measure $\mu^\infty(\cdot\,|s,(\theta,x^t,y^t),\xi)$ over $\Theta \times X^\infty \times Y^\infty$ much as in Lemma A.2: begin with the probability measure $\delta_\theta \otimes \delta_{x^t} \otimes \xi \otimes \delta_{y^t}$ over $\Theta \times X^{t+1} \times Y^t$; use $\alpha(\cdot\,|x_{t+1},\theta)$ to move from

$\Theta \times X^{t+1} \times Y^t$ to $\Theta \times X^{t+1} \times Y^{t+1}$; then use s exactly as in the proof of Lemma A.2. Thirdly, let

$$v_s(\xi \mid \theta, x^t, y^t) = \int (1-\delta) \sum_{u=1}^{\infty} \delta^{u-1} B(\hat{x}_{t+u}, \hat{\theta}) d\mu^{\infty}(\hat{\theta}, \hat{x}^{\infty}, \hat{y}^{\infty} \mid s, (\theta, x^t, y^t), \xi)$$

be the continuation payoff obtained when the parameter is θ, x^t is the past history of actions, y^t is the past history of observations, and ξ is the probability distribution over actions employed in period $t+1$. We shall also need the following obvious facts. First, $\mu^t(\cdot \mid s, Q)$ depends on s only insofar as it depends on s^t. Secondly, $\mu^{\infty}(\cdot \mid s, (\theta, x^t, y^t), \xi)$ and $v_s(\xi \mid \theta, x^t, y^t)$ depend on s only insofar as they depend on ${}^{t+2}s$. Thirdly, $v_s(\xi \mid \theta, x^t, y^t)$ is bounded and lower semicontinuous in ξ.

Turning to the proof itself, let s be any strategy. Because b is bounded, there exists T such that

$$v_{\left[s^T, {}^{T+1}\bar{s}\right]}(Q) > v_s(Q) - \epsilon$$

for all continuous ${}^{T+1}\bar{s}$. Next, from Lemma A.1 it follows that we can find strategies $\bar{s}_{Tn}: Y^{T-1} \to \mathcal{PM}(X)$ such that, regarded as a random variable on $\Theta \times X^{T-1} \times Y^{T-1}$, $\bar{s}_{Tn} \to s_T$ a.s. relative to the measure $\mu^{T-1}(\cdot \mid s^{T-1}, Q)$. But

$$v_{\left[s^{T-1}, \bar{s}_{Tn}, {}^{T+1}\bar{s}\right]}(Q) = \int (1-\delta) \left[\sum_{t=1}^{T-1} \delta^{t-1} B(x_t, \theta) \right] d\mu^{T-1}(\theta, x^{T-1}, y^{T-1} \mid s^{T-1}, Q)$$

$$+ \int \left[\int (1-\delta) \delta^{T-1} B(x_T, \theta) d\bar{s}_{Tn}(x_T \mid y^{T-1}) \right] d\mu^{T-1}(\theta, x^{T-1}, y^{T-1} \mid s^{T-1}, Q)$$

$$+ \int v_{[s^{T},T+1,\bar{s}]}(\bar{s}_{Tn}(\cdot|y^{T-1}),\theta,x^{T-1},y^{T-1})d\mu^{T-1}(\theta,x^{T-1},y^{T-1}|s^{T-1},Q),$$

with a similar formula for $v_{[s^{T-1},s_T,T+1,\bar{s}]}$. But the first term on the RHS is independent of n. In the second, $B(x_T,\theta)$ is lsc in x_T and $\bar{s}_{Tn}(\cdot|y^{T-1})$ converges weakly to $s_T(\cdot|y^{T-1})$ for almost all y^{T-1}. In the third, $v_{[s^{T},T+1,\bar{s}]}$ is lsc in $\bar{s}_{Tn}(\cdot|y^{T-1})$, and $\bar{s}_{Tn}(\cdot|y^{T-1})$ converges weakly to $s_T(\cdot|y^{T-1})$ as before. It follows that

$$\liminf_{n\to\infty} v_{[s^{T-1},\bar{s}_{Tn},T+1,\bar{s}]}(Q) \geq v_{[s^{T-1},s_T,T+1,\bar{s}]}(Q),$$

and therefore that there exists N such that

$$v_{[s^{T-1},\bar{s}_{TN},T+1,\bar{s}]}(Q) > v_{[s^{T-1},s_T,T+1,\bar{s}]}(Q) - \epsilon/T.$$

Let $\bar{s}_{tT} = \bar{s}_{TN}$.

Iterating this argument we eventually obtain a strategy $\bar{s}$ such that $v_{\bar{s}}(Q) > v_s(Q) - 2\epsilon$. This completes the proof of the lemma. □

Proof of Theorem A.5 Define $\underline{v}(Q) = \sup_{\bar{s}} v_{\bar{s}}(Q)$, where $\bar{s}$ varies over all continuous strategies, and $\bar{v}(Q) = \sup_s v_s(Q)$, where s varies over all strategies. Clearly $\underline{v}$ and $\bar{v}$ are bounded, and $\underline{v} \leq \bar{v}$. By Lemma A.3, $\underline{v} \geq \bar{v}$, so in fact $\underline{v} \equiv \bar{v}$. By Lemma A.2, each $v_{\bar{s}}$ is lower semicontinuous, so $\bar{v}$ too is lower semicontinuous. It remains to relate $\bar{v}$ to v. (The difference between $\bar{v}$ and v is that $\bar{v}$ is defined for general strategies taking on random values, whereas v is defined only for strategies taking on deterministic values.) Certainly $\bar{v} \geq v$, since v is the supremum over a smaller set of strategies. On the other hand, standard

considerations show that, for any given strategy s with possibly randomised values, there exists a deterministic $\hat{s}$ that does at least as well. So $\bar{v} \leq v$. □

We turn now to two important corollaries of Theorem A.5.

Corollary A.1 Suppose that the conditions of Theorem A.5 hold. Then (B) holds.

Proof Consider the epigraph $E = \{(Q,w) \mid w \geq v(Q)\}$ of v. For each $(\overline{Q},\overline{w}) \in (\mathcal{PM}(\Theta) \times \mathbb{R})\backslash E$, i.e. for which $v(\overline{Q}) > \overline{w}$, there exists a continuous strategy s such that $v_s(\overline{Q}) > \overline{w}$. Let $A(s) = \{(Q,w) \mid v_s(Q) > w\}$. Because v_s is lower semicontinuous, A(s) is open. By choice of s, $(\overline{Q},\overline{w}) \in A(s)$. So the sets A(s) obtained as $(\overline{Q},\overline{w})$ varies over $(\mathcal{PM}(\Theta) \times \mathbb{R})\backslash E$ from an open cover for this set. Because this set is separable, we may select a countable subcover $\{A(s) \mid s \in S\}$. The set of strategies S then serves in the capacity required by (B). □

Corollary A.2 Suppose that the conditions of Theorem A.5 hold. Then $v(Q_t) \to v(Q_\infty)$ and $m(Q_t) \to m(Q_\infty)$.

Proof Consider the case of m. Because m is lower semicontinuous and convex, and because $\mathcal{PM}(\Theta) \times \mathbb{R}$ is separable, m can be expressed as the upper envelope of a countable number of continuous linear functionals. Jensen's inequality therefore applies, and $\{m(Q_t) \mid 1 \leq t \leq \infty\}$ is a bounded submartingale. Hence, by the submartingale convergence theorem, there exists $\bar{M}$ such that $m(Q_t) \to \bar{M}$ a.s. as $t \to \infty-$. Since $m(Q_t) \leq E[m(Q_\infty) \mid \mathcal{F}_t]$ a.s. and $E[m(Q_\infty) \mid \mathcal{F}_t] \to m(Q_\infty)$ a.s., $\bar{M} \leq m(Q_\infty)$ a.s. On the other hand, $\bar{M} = \lim m(Q_t) \geq m(Q_\infty)$ a.s. by the lower semicontinuity of m. □

Theorem A.6 Suppose that s is optimal, that a is continuous in (x,θ), and that b is continuous in (x,θ). Then, with probability one, all limit points of $\{x_t \mid 1 \leq t \leq \infty\}$ maximise $B(\cdot,Q_\infty)$.

Proof Re–examining the proof of Theorem A.4, we see that it incidentally shows that $m(Q_t) - B(x_{t+1},Q_t) \to 0$ a.s. Restrict attention to states of the world in which this occurs, and in which moreover $Q_t \to Q_\infty$ and $m(Q_t) \to m(Q_\infty)$. Let x_∞ by any limit point of $\{x_t \mid 1 \leq t < \infty\}$ in such a state of the world. Moving to a subsequence if necessary, we may assume that $x_t \to x_\infty$. We have

$$B(x_\infty,Q_\infty) \geq \limsup_{t\to\infty} B(x_{t+1},Q_t)$$

(by upper semicontinuity of B)

$$= m(Q_\infty)$$

(by our choice of state). That is, x_∞ maximises $B(\cdot,Q_\infty)$. □

Corollary A.3 Suppose that the conditions of Theorem A.6 are satisfied, and that adequate learning occurs. Then, with probability one, every limit point of $\{x_t \mid 1 \leq t < \infty\}$ maximises $B(\cdot,\theta)$.

Proof This can be proved in almost exactly the same way as Theorem A.6. We know that, with probability one: $m(Q_t) - B(x_{t+1},Q_t) \to 0$; $Q_t \to Q_\infty$; and $m(Q_t) \to M(\theta)$. For states of the world for which this is the case, $B(x_\infty,Q_\infty) \geq M(\theta)$ for all limit points x_∞ of $\{x_t \mid 1 \leq t \leq \infty\}$. □

We turn now to the proof of our existence theorem. We will accomplish this proof under the assumptions (E1)–(E4). It should be noted that these assumptions are unambiguously stronger than (A1), (A2′) and (A3′). The crucial step is to show that Bayes' Law is continuous in an appropriate sense. This is the content of Lemma A.4.

In order to state Lemma A.4 we need some notation. First, standard considerations show that there is a Borel measurable mapping q: $\mathcal{PM}(\Theta) \times X \times Y \to \mathcal{PM}(\Theta)$ such that, for all Q, x, and y, q(Q,x,y) is the agent's posterior when his priors are Q, he takes action x and he observes signal y. Secondly, because $X \times \Theta$ is separable and F is norm continuous in (x,θ), there exists a probability measure λ on Y and a Borel measurable mapping $f(\cdot\,|\,\cdot,\cdot): Y \times X \times \Theta \to [0,\infty)$ such that $f(\cdot\,|\,x,\theta) \in L^1(\lambda)$ and $dF(\cdot\,|\,x,\theta) = f(\cdot\,|\,x,\theta)d\lambda$ for all (x,θ). Moreover $f(\cdot\,|\,x,\theta)$, regarded as an element of $L^1(\lambda)$, varies continuously with (x,θ) in the usual L^1 norm.

Lemma A.4 Suppose that (E1) and (E2) hold. Let $\{(Q_n,x_n)\,|\,n \geq 1\}$ be a sequence that converges to (Q,x). Then $q(Q_n,x_n,\cdot)$, regarded as a random variable on Y, converges in $\int F(\cdot\,|\,x,\theta)dQ(\theta)$ – probability to $q(Q,x,\cdot)$, similarly regarded.

Note that $\int F(\cdot\,|\,x,\theta)dQ(\theta)$ is simply the marginal distribution of the observation when the agent has priors Q and chooses action x.

Proof It suffices to prove that $\int gd[q(Q_n,x_n,\cdot)]$, regarded as a real–valued random variable on Y, converges in $\int F(\cdot\,|\,x,\theta)dQ(\theta)$ – probability to $\int gd[q(Q,x,\cdot)]$, similarly regarded, for any given bounded continuous $g: \Theta \to \mathbb{R}$. To this end, let g be such a function. Because the observation distributions are absolutely continuous with respect to λ, we have

$$\int gd[q(Q_n,x_n,y)] = \frac{\int g(\varphi)f(y\,|\,x_n,\varphi)\,dQ_n(\varphi)}{\int f(y\,|\,x_n,\varphi)dQ_n(\varphi)}$$

for all y in a set of $\int F(\cdot\,|x_n,\varphi)dQ_n(\varphi)$ – probability 1.

Since $Q_n \to Q$ we may find random variables Φ_n and Φ on some auxiliary probability space such that Φ_n has distribution Q_n, Φ has distribution Q, and $\Phi_n \to \Phi$ a.s. We then have

$$\int |\int g(\varphi)f(y|x_n,\varphi)dQ_n(\varphi) - \int g(\varphi)f(y|x,\varphi)dQ\,(\varphi)|\,d\lambda(y)$$

$$= \int |E[g(\Phi_n)f(y|x_n,\Phi_n)] - E[g(\Phi)f(y|x,\Phi)]|\,d\lambda(y)$$

$$\leq \|g\|_\infty E[\int |f(y|x_n,\Phi_n) - f(y|x,\Phi)|\,d\lambda(y)].$$

But $\int |f(y|x_n,\Phi_n) - f(y|x,\Phi)|\,d\lambda(y) \leq 2$, and converges to zero a.s. So, applying Lebesgue's Bounded Convergence Theorem, we conclude that $\int g(\varphi)f(\cdot\,|x_n,\varphi)dQ_n(\varphi)$, regarded as an element of $L^1(\lambda)$, converges in norm to $\int g(\varphi)f(\cdot\,|x,\varphi)dQ(\varphi)$, similarly regarded. This implies in particular that $\int g(\varphi)f(\cdot\,|x_n,\varphi)dQ_n(\varphi)$ converges in λ–probability to $\int g(\varphi)f(\cdot\,|x,\varphi)dQ(\varphi)$. Similarly, $\int f(\cdot\,|x_n,\varphi)dQ_n(\varphi)$ converges in λ–probability to $\int f(\cdot\,|x,\varphi)dQ(\varphi)$. Hence, finally, $\int gd[q(Q_n,x_n,\cdot)]$ converges in λ-probability to $\int gd[q(Q,x,\cdot)]$ on the set of y such that $\int f(y|x,\varphi)dQ(\varphi) > 0$. This is the required conclusion. □

Theorem A.7 Suppose that (E1)–(E4) hold. Then the agent possesses an optimal strategy.

Proof The proof reduces to solving the Bellman equation for our problem, and showing that it can be used to construct an optimal strategy. Although we merely sketch it, we hope that our sketch will illustrate the role played by the main assumptions.

Consider the mapping Φ defined on the Banach space of bounded, Borel measurable functions on $\mathcal{PM}(\Theta)$ by

$$[\Phi(w)](Q) = \sup_x\{(1-\delta)\int b(x,\theta,z)d(Q\otimes R)(\theta,z) + \delta\int w(q(Q,x,y))f(y|x,\theta)d(Q\otimes\lambda)(\theta,y)\}.$$

It is easily checked that Φ is a contraction of order $\delta < 1$, and that it is monotonic in the sense that $\Phi(w_2) \leq \Phi(w_1)$ if $w_2 \leq w_1$. Finally, it follows from the facts that b is usc in x and that q is continuous in (Q,x), that Φ maps usc functions to usc functions.

Let K be a bound for b, set $w_0 \equiv K$, and set $w_{n+1} = \Phi(w_n)$ for all $n \geq 0$. Because Φ is a contraction, w_n converges uniformly to the unique fixed point w_∞ of Φ. Next, it is clear that $w_1 \leq w_0$. It then follows by induction that w_n converges monotonically. Hence $w_\infty = \inf_n w_n$. Finally, w_0 is clearly usc. Hence all the w_n are usc, and w_∞ is usc as an infimum of usc functions. Because w_∞ is usc,

$$\operatorname{argmax}_x\{(1-\delta)\int b(x,\theta,z)d(Q\otimes R)(\theta,z) + \delta\int w_\infty(q(Q,x,y))f(y|x,\theta)d(Q\otimes\lambda)(\theta,y)\}$$

is a non–empty and compact–valued correspondence defined on $\mathcal{PM}(\Theta)$. It is also measurable in an appropriate sense. It therefore admits a measurable selection x^*: $\mathcal{PM}(\Theta) \to X$. Standard considerations then show that w_∞ is the payoff obtained when x^* is employed, and that any other strategy yields a payoff of at most w_∞. It follows that w_∞ is the value function for our problem, and that x^* is an optimal strategy. □

It should be noted that, as the proof of Theorem A.7 makes clear, (E1)–(E4) represent a set of conditions alternative to (A1)–(A3) under which dynamic programming can be justified in our model. In particular, under (E1)–(E4) one can prove an analogue of Theorem A.4 on the exhaustion of learning opportunities in the long run. This should help

to explain the special prominence which we gave to that result. One cannot, however, prove continuity results like Corollary A.2, on the continuity of $v(Q_t)$ and $m(Q_t)$ in t, or even like Theorem A.6, on the continuity of optimal actions x_t in t.

We illustrate these points by means of two examples which should, incidentally, highlight the role played by the various continuity assumptions that we have made. In the first example: $\Theta = X = [0,1]$, $Z = \mathbb{R}$, $a(x,\theta,z) = \theta + z$, $b(x,\theta) = 1$ if $x = \theta$ and 0 otherwise, Q is the uniform distribution over [0,1], and R is the standard normal distribution. This example satisfies (E1)–(E4), but fails to satisfy (A3). In it, learning is purely passive, and the agent's only objective is therefore to maximise $B(\cdot,Q_t)$ in every period t + 1. $B(\cdot,Q_t)$ is, however, identically zero. For the agent's beliefs about θ have a density relative to Lebesgue measure on [0,1]. So he is indifferent between all choices of action x, and $m(Q_t) = 0$. On the other hand, $Q_\infty(\cdot \mid \omega) = \delta_\theta$ with probability one. So $B(x,Q_\infty) = 1$ if $x = \theta$ and 0 otherwise, the agent has the unique optimal action $x = \theta$ in the asymptotic problem, and $m(Q_\infty) = 1$. Hence, with probability one, $m(Q_t)$ fails to converge. Moreover there exist optimal strategies s such that, with probability one, no limit point of $\{x_t \mid 1 \leq t\}$ maximises $B(\cdot,Q_\infty)$. This latter finding is, however, pathological. There does exist an optimal strategy s for which, with probability one, every limit point of $\{x_t \mid 1 \leq t\}$ maximises $B(\cdot,Q_\infty)$. So the feeling that Theorem A.6 relies, in some sense, on upper rather than lower semicontinuity is vindicated to some extent.

In the second example: $\Theta = X = [0,1]^2$, $Z = \mathbb{R}$, $a(x,\theta,z) = (\theta_1 + z, \theta_2)$ if $x_1 = \theta_1$ and $(\theta_1 + z, 0)$ if $x_1 \neq \theta_1$, $b(x,\theta,z) = -(x_2 - \theta_2)^2$, Q is the uniform distribution over $[0,1]^2$, and R is the standard normal distribution. This example satisfies (A1) and (A3) but not (A2). More specifically, b is continuous but a is not. It can be analysed in much the same way as the previous example. The crucial point is that the agent wants to learn θ_2, so he can set $x_2 = \theta_2$. However, to do so he must set $x_1 = \theta_1$. This he cannot achieve because he learns about θ_1 only slowly and passively. Since b is continuous, $m(Q_t) \to m(Q_\infty)$ a.s.

However $v(Q_t) = 0$ a.s. whereas $v(Q_\infty) > 0$ a.s. In other words, new learning possibilities spring up in the limit.

We conclude this appendix with proofs of Theorems 3.1, 3.3, 4.2, and 4.3.

Proof of Theorem 3.1

A detailed proof conceals more than it reveals, so we merely sketch the main steps. Also, we have already remarked that it is sufficient to consider the case $m(Q) = v(Q)$, so let us assume this. Let x^* maximise $B(\cdot,Q)$ and assume, for a contradiction, that $m(Q) < E(M(\theta))$.

The main obstacle that needs to be overcome is that of constructing a global estimate of $b(\cdot,\theta)$ given approximate knowledge of its derivatives at x^*. More precisely, for all n and $\alpha_0,\alpha_1,\ldots,\alpha_n 2$, we need to construct an estimate $e(\cdot;n, <\alpha_i | 0 \le i \le n^2>)$ of $b(\cdot,\theta)$ with the following properties: (i) if $<\alpha_i | i \ge 0>$ are the true derivatives of b at x^*, then $e \to b(\cdot,\theta)$ as $n \to \infty$; (ii) for any given n, e is continuous in $<\alpha_i | 0 \le i \le n^2>$; (iii) e is bounded. (Here the range of e, which consists of continuous functions, is endowed with the topology of uniform convergence.)

Suppose, for the moment, that such an estimate e is given, and let $\beta = (E[M(\theta)] - m(Q))/3$. Then, for suitable choice of n and ϵ, the following scheme yields an improvement in the agent's payoff. Over the first $n^2 + 1$ periods, try $n^2 + 1$ actions that are evenly spaced with spacing ϵ, ensuring that one of these actions is x^*. Next, use a suitable differencing scheme to arrive at an estimate $<\alpha_i | 0 \le i \le n^2>$ of the first $n^2 + 1$ derivatives of $b(\cdot,\theta)$ at x^*. (This estimate will depend on n, ϵ and the $n^2 + 1$ observations obtained.) Finally, choose x to maximise the associated estimate $e(\cdot;n, <\alpha_i | 0 \le i \le n^2>)$ of the payoff function.

The choice of n and ϵ can be made as follows. By the first property of e,

$$e(\cdot;n,<\frac{\partial^i b}{\partial x^i}(x^*,\theta)\,|\,0\le i\le n^2>)\to b(\cdot,\theta)$$

uniformly for <u>every</u> θ (but not, of course, uniformly in θ). Hence, by property (iii), we may find $\bar{n}$ such that

$$E\left[\|e(\cdot;\bar{n},<\frac{\partial^i b}{\partial x^i}(x^*,\theta)>)-b(\cdot,\theta)\|_\infty\right]<\beta. \qquad ...(A.1)$$

Next, by choosing $\bar{\epsilon}=\epsilon(\bar{n})$ sufficiently small, we can simultaneously ensure that

$$E\left[\|e(\cdot;\bar{n},<\alpha_i>)-e(\cdot;\bar{n},<\frac{\partial^i b}{\partial x^i}(x^*,\theta)>)\|_\infty\right]<\beta \qquad ...(A.2)$$

and that the cost of experimentation over the first $\bar{n}^2+1$ periods is less than $\delta^{n^2+1}\beta$. Finally, in view of (A.1) and (A.2), playing x from period $n^{-2}+2$ on yields an expected payoff of more than $m(Q)+\beta$ per period. This increased payoff more than compensates the cost of experimentation.

It remains to construct the estimate e. This would present no problem if we knew that $b(\cdot,\theta)$ could be expanded globally as a power series about x^*. For then we could simply set

$$\bar{e}(x;n,<\alpha_i>)=\sum_{i=0}^{n^2}\alpha_i(x-x^*)^i$$

and $e = \max\{-K,\min\{K,\bar{e}\}\}$, where K is a bound for b. But we must contend with the fact that b can only be expanded locally about any given point.

We cope with this difficulty as follows. Let $\ell = (\bar{x} - \underline{x})/n$. Approximate $b(\cdot,\theta)$ in the interval $[x^*,x^* + \ell]$ by

$$\sum_{i=0}^{n^2} \alpha_i (x - x^*)^i,$$

and approximate the first $(n^2 + 1) - n$ derivatives of $b(\cdot,\theta)$ by the first $(n^2 + 1) - n$ derivatives of this polynomial. Next, let $<\beta_i \mid 0 \leq i \leq n^2 - n>$ be the values of these derivatives at $x^* + \ell$. Approximate $b(\cdot,\theta)$ in the interval $[x^* + \ell, x^* + 2\ell]$ by

$$\sum_{i=0}^{n^2} \beta_i (x - x^* - \ell)^i,$$

and approximate the first $(n^2 + 1) - 2n$ derivatives of $b(\cdot,\theta)$ by the first $(n^2 + 1) - 2n$ derivatives of this polynomial. Proceeding in this way, one reaches $\bar{x}$ after at most n steps, at which point one has an estimate of at least the first $(n^2 + 1) - n^2 = 1$ derivatives of $b(\cdot,\theta)$, i.e. of $b(\bar{x},\theta)$. Similarly, proceeding leftwards instead of rightwards, one can estimate $b(\cdot,\theta)$ up to $\underline{x}$.

This procedure works for the following reasons. First, since $b(\cdot,\theta)$ is real analytic, we know that there exists $r(\theta) > 0$ such that the radius of convergence of the power series for $b(\cdot,\theta)$ about x is at least $r(\theta)$ for all $x \in [\underline{x},\bar{x}]$. Hence, for n sufficiently large, our step size ℓ is less than $r(\theta)$ and approximation by polynomial is valid. Secondly, by only estimating the first $(n^2 + 1) - n$ derivatives of $b(\cdot,\theta)$ at $x^* + \ell$, we ensure that we have at

least n terms in our approximations. These approximations therefore improve as n gets large. □

Proof of Theorem 3.3

The first step is to tackle the learning problem in isolation. As remarked in the text, we must ensure that the agent's estimates of the $B(\xi_n,\theta)$ are sufficiently accurate. To this end, note that for each n and θ, $\frac{1}{T}\Sigma_{t=1}^{T} b(\xi_n,\theta,z_t) \to B(\xi_n,\theta)$ outside a null set of Z^∞. Hence $\frac{1}{T}\Sigma_{t=1}^{T} b(\xi_n,\theta,z_t) \to B(\xi_n,\theta)$ P–a.s. Hence, for all N, we can find T_N such that

$$\text{prob}\{\sup_{n\le N} |\tfrac{1}{T}\Sigma_{t=1}^{T} b(\xi_n,\theta,z_t) - B(\xi_n,\theta)| \ge \tfrac{1}{N}\} \le N^{-2} \qquad \ldots(A.3)$$

for all $T \ge T_N$.

The second step exploits the sequence $\{T_N\}$ to construct an optimal strategy. In this strategy, information–accumulation periods lasting β_N stages alternate with payoff–accumulation periods lasting τ_N stages. The first information-accumulation period consists of T_1 trials of ξ_1. For $N \ge 2$, the N^{th} information-accumulation period consists of $(T_N - T_{N-1})$ trials of ξ_n for all $1 \le n \le N-1$, and of T_N trials of ξ_N. Hence, at the outset of the N^{th} payoff–accumulation period, the agent has a total of T_N observations on each ξ_n with $1 \le n \le N$. Assume that he estimates $B(\xi_n,\theta)$ by averaging these observations, and let $\xi_{i(N)}$ yield the highest estimate. Then the N^{th} payoff–accumulation period consists in playing $\xi_{i(N)}$ a total of τ_N times. Let α_N be the total number of stages before the N^{th} payoff–accumulation period starts. Then $\alpha_{N+1} = \alpha_N + \tau_N + \beta_{N+1}$, where $\beta_N = NT_N - (N-1)T_{N-1}$ is the number of stages of the N^{th} information–accumulation period. We choose α_N, τ_N so that $\tau_N/\alpha_{N+1} \to 1$ as $N \to \infty$. Thus, for large N, the payoffs accumulated in the N^{th} payoff–accumulation period outweigh all previous payoffs

(including those from payoff–accumulation periods) and also the payoffs obtained in the subsequent information–accumulation period.

The third and final step verifies that this strategy is indeed optimal. It suffices to check that $E(\liminf_{T\to\infty} \frac{1}{T}\Sigma_{t=1}^{T}\pi_t) \geq E(M(\theta))$. The most difficult case is clearly that in which the $(N+1)^{th}$ information accumulation period has just finished, so we confine ourselves to the case $T = \alpha_{N+1}$. Let K be a bound for b. Then $\pi_t \geq -K$ for the first α_N and the last β_{N+1} stages.

Hence

$$\frac{1}{\alpha_{N+1}} \sum_{t=1}^{\alpha_{N+1}} \pi_t \geq -\frac{K}{\alpha_{N+1}}\left[\alpha_{N+1} - \tau_N\right] + \frac{\tau_N}{\alpha_{N+1}}\frac{1}{\tau_N} \sum_{t=\alpha_N+1}^{\alpha_N+\tau_N} b(\xi_{i(N)},\theta,z_t).$$

The first term on the RHS is easily dealt with. It converges to zero by choice of τ_N. The second requires more care. Note first that certainly

$$\sup_{n\leq N} \left| \left[\frac{1}{\tau_N} \sum_{t=\alpha_N+1}^{\alpha_N+\tau_N} b(\xi_n,\theta,z_t)\right] - B(\xi_n,\theta)\right| \to 0 \text{ a.s.}$$

as $N \to \infty$. This follows from (A.3) above, the Borel–Cantelli lemma, and the fact that $\tau_N \geq T_N$. It implies that

$$\left| \left[\frac{1}{\tau_N} \sum_{t=\alpha_N+1}^{\alpha_N+\tau_N} b(\xi_{i(N)},\theta,z_t)\right] - B(\xi_{i(N)},\theta)\right| \to 0 \text{ a.s.}$$

So in order to tie down the behaviour of the second term it suffices to tie down the behaviour of $B(\xi_{i(N)},\theta)$. But by definition of i(N),

$$\max_{n\leq N}\frac{1}{T_N}\sum_{t\in I(n,N)} b(\xi_n,\theta,z_t) = \frac{1}{T_N}\sum_{t\in I(i(N),N)} b(\xi_{i(N)},\theta,z_t)$$

where I(n,N) is the set of periods in which the first T_N observations on ξ_n are made. Also, by (A.3) and the Borel–Cantelli lemma again, we have

$$\left[\max_{n\leq N}\frac{1}{T_N}\sum_{t\in I(n,N)} b(\xi_n,\theta,z_t)\right] - M_N(\theta) \to 0 \text{ a.s.}$$

and

$$\left[\frac{1}{T_N}\sum_{t\in I(n,N)} b(\xi_{i(N)},\theta,z_t)\right] - B(\xi_{i(N)},\theta) \to 0 \text{ a.s.}$$

So $B(\xi_{i(N)},\theta) \to M(\theta)$ a.s. The result follows. □

Proof of Theorem 4.2

The theorem is obvious if X is a degenerate interval, so we assume that $X = [\underline{x}, \overline{x}]$ with $\overline{x} > \underline{x}$. Also, it suffices to verify the theorem for sequences $\{R_n\}$ such that $R_n \to \delta_0$. So we fix attention on a particular such sequence.

For any given n we know from Lemma A.5 that the agent's posterior converges. In a convenient abuse of notation, we donate the limiting posteriors by Q_n. (Note that Q_n should be thought of as a random variable defined on the measure space $(\Omega,\mathcal{F})$ and taking values in $M(\Theta)$.) We also know that the asymptotic payoff is $m(Q_n)$. So we may define a

random variable ξ_n such that $\xi_n(\omega)$ maximises $B(\cdot, Q_n(\omega))$ for each $\omega \in \Omega$. Then, in the game with prior $Q_n(\omega)$ (and noise distribution R_n), playing $\xi_n(\omega)$ forever is an optimal strategy.

Now suppose that $\alpha > 0$ and $\beta > 0$ are given. Then, if $\xi_n(\omega) < \bar{x}$, we may pick $0 < \gamma(\omega) \leq \beta$ such that $\xi_n(\omega) + \alpha\gamma(\omega) < \bar{x}$. Since playing $\xi_n(\omega)$ forever is optimal, it is in particular superior to the alternative strategy: (i) play $\xi_n(\omega)$ in stage 1; (ii) play $\xi_n(\omega) + \gamma(\omega)$ in stage 2; (iii) if the payoff in stage 2 exceeds that in stage 1 then play $\xi_n(\omega) + \alpha\gamma(\omega)$ for evermore. This implies that

$$0 \geq (1-\delta)\int\Big[b(\xi_n + \gamma, \theta, z_2) - b(\xi_n, \theta, z_2)\Big] d(Q_n \circledast R_n)(\theta, z_2)$$

$$+ \delta\int \chi[b(\xi_n+\gamma,\theta,z_2) - b(\xi_n,\theta,z_1)]\Big[b(\xi_n+\alpha\gamma,\theta,z_3) - b(\xi_n,\theta,z_3)\Big] d(Q_n \circledast R_n^3)(\theta, z_1, z_2, z_3),$$

where χ is the indicator function for $(0,\infty)$, and where we have suppressed the dependence of Q_n, ξ_n and γ on ω. Let $\tilde{\xi}_n \in (\xi_n, \xi_n + \gamma)$ and $\hat{\xi}_n \in (\xi_n, \xi_n + \alpha\gamma)$ be chosen such that $b(\xi_n + \gamma, \theta, z_2) - b(\xi_n, \theta, z_2) = \gamma(\partial b/\partial x)(\tilde{\xi}_n, \theta, z_2)$ and $b(\xi_n + \alpha\gamma, \theta, z_3) - b(\xi_n \theta, z_3)$ $= \alpha\gamma(\partial b/\partial x)(\hat{\xi}_n, \theta, z_3)$. Then the above inequality reduces to

$$0 \geq (1-\delta)\int \frac{\partial b}{\partial x}(\tilde{\xi}_n, \theta, z_2) d(Q_n \circledast R_n)$$

$$+ \alpha\delta\int \chi[b(\xi_n + \gamma, \theta, z_2) - b(\xi_n, \theta, z_1)] \frac{\partial b}{\partial x}(\hat{\xi}_n, \theta, z_3) d(Q_n \circledast R_n^3) \qquad \ldots(A.4)$$

where we have suppressed the arguments of the measures. The bulk of the remainder of the proof consists in showing that, for suitably chosen α, β and γ, this inequality implies that

$$\int \chi[\bar{x} - \hat{\xi}_n] \max\{0, \frac{\partial b}{\partial x}(\hat{\xi}_n, \theta, 0)\} dP_n(\omega, \theta) \qquad \ldots(A.5)$$

converges to zero as $n \to \infty$, where P_n is the probability measure on $\Omega \times \Theta$ obtained by combining the probability measure P on Ω with the transition probability Q_n.

As a first step in this direction, we need to complete the definition of $\tilde{\xi}_n$, $\hat{\xi}_n$ and γ. Set $\tilde{\xi}_n = \bar{x}$, $\hat{\xi}_n = \bar{x}$ and $\gamma = 0$ when $\xi_n = \bar{x}$. We also need a version of the inequality (A.4) for this case. The trivial inequality $0 \geq 0$ will suffice for our purposes. We can now integrate with respect to ω to obtain

$$0 \geq \frac{1-\delta}{\alpha\delta}\int \chi[\bar{x} - \hat{\xi}_n]\frac{\partial b}{\partial x}(\tilde{\xi}_n,\theta,z_2)d(P_n \circledast R_n)$$

$$+ \int \chi[b(\xi_n + \gamma,\theta,z_2) - b(\xi_n,\theta,z_1)]\left[\frac{\partial b}{\partial x}(\hat{\xi}_n,\theta,z_3) - \frac{\partial b}{\partial x}(\hat{\xi}_n,\theta,0)\right]d(P_n \circledast R_n^3)$$

$$+ \int\left[\chi[b(\xi_n+\gamma,\theta,z_2) - b(\xi_n,\theta,z_1)] - \chi[b(\xi_n+\gamma,\theta,0) - b(\xi_n,\theta,0)]\right]\frac{\partial b}{\partial x}(\hat{\xi}_n,\theta,0)d(P_n \circledast R_n^3)$$

$$+ \int\left[\chi[b_n(\xi_n + \gamma,\theta,0) - b(\xi_n,\theta,0)] - \chi\left[\frac{\partial b}{\partial x}(\hat{\xi}_n,\theta,0)\right]\right]\frac{\partial b}{\partial x}(\hat{\xi}_n,\theta,0)\, dP_n$$

$$+ \int \chi[\bar{x} - \hat{\xi}_n]\max\{0,\frac{\partial b}{\partial x}(\hat{\xi}_n,\theta,0)\}\, dP_n, \qquad \ldots(A.6)$$

where we have again suppressed the arguments of the measures. (Note that all five integrands are zero when $\xi_n = \bar{x}$.)

The next step is to estimate the terms on the right–hand side of this inequality. For this purpose we introduce some further notation. Let ψ be the indicator function of the set $\{0\}$. Let

$$c_1(\alpha,\beta,\theta) = \sup_{x_1,x_2,x_3}\left|\frac{\partial b}{\partial x}(x_3,\theta,0)\left[\chi[b(x_2,\theta,0) - b(x_1,\theta,0)] - \chi\left[\frac{\partial b}{\partial x}(x_3,\theta,0)\right]\right]\right|,$$

where $x_1,x_2,x_3 \in X$, $x_2 \leq x_1 + \beta$, and $x_3 \leq x_1 + \alpha\beta$. Let

$$c_2(\alpha,\beta,\theta) = \sup_{x_1,x_2,x_3} \left|\frac{\partial b}{\partial x}(x_3,\theta,0)\psi[b(x_2,\theta,0) - b(x_1,\theta,0)]\right|,$$

where once again $x_1,x_2,x_3 \in X$, $x_2 \leq x_1 + \beta$, and $x_3 \leq x_1 + \alpha\beta$. Let

$$d_1(z,\theta) = \sup_x \left|\frac{\partial b}{\partial x}(x,\theta,z) - \frac{\partial b}{\partial x}(x,\theta,0)\right|,$$

where $x \in X$. Finally, let

$$d_2(z_1,z_2,z_3,\alpha,\beta,\theta) = \sup_{x_1,x_2,x_3} \left|\frac{\partial b}{\partial x}(x_3,\theta,0)\Big[\chi[b(x_2,\theta,z_2) - b(x_1,\theta,z_1)]\right.$$

$$\left. - \chi[b(x_2,\theta,0) - b(x_1,\theta,0)]\Big]\right|,$$

where $x_1,x_2,x_3 \in X$, $x_2 \leq x_1 + \beta$, and $x_3 \leq x_1 + \alpha\beta$. Then it is easily checked that all four of these functions are dominated by $D(\theta)$.

We can now proceed to the estimation itself. Suppose that $\epsilon > 0$ is given. Pick $\alpha = \alpha(\epsilon)$ sufficiently large that $(1 - \delta)(\int D(\theta)dQ(\theta))/\alpha\delta < \epsilon$. Next, note that $c_1(\alpha,\beta,\theta) \to 0$ and $c_2(\alpha,\beta,\theta) \to 0$ as $\beta \to 0$. Pick $\beta = \beta(\epsilon,\alpha)$ sufficiently small that $\int c_1(\alpha,\beta,\theta)dQ(\theta) < \epsilon$, $\int c_2(\alpha,\beta,\theta)dQ(\theta) < \epsilon$, and $\beta < \epsilon$. Thirdly, note that $d_1(z,\theta) \to 0$ as $z \to 0$ and that

$$\limsup_{(z_1,z_2,z_3)\to 0} d_2(z_1,z_2,z_3,\alpha,\beta,\theta) \leq c_2(\alpha,\beta,\theta).$$

We can therefore pick $N = N(\epsilon,\alpha,\beta)$ sufficiently large that $\int d_1(\theta,z)d(Q \times R_n) < \epsilon$, and

$$\int d_2(z_1,z_2,z_3,\alpha,\beta,\theta)d(Q \times R_n^3) < \epsilon + \int c_2(\alpha,\beta,\theta)dQ(\theta)$$

for all $n \geq N$. Now suppose that α, β and N are chosen in this way. Then the choice of α ensures that the first term on the right–hand side of (A.6) is smaller than ϵ in absolute value; the choice of β ensures that the same is true of the fourth term; the choice of N ensures that the second term is less than ϵ in absolute value, and that the third is less than $\epsilon + \int c_2(\alpha,\beta,\theta)dQ(\theta)$ in absolute value; and the choice of β ensures that $\int c_2(\alpha,\beta,\theta)dQ(\theta) < \epsilon$. So, overall, (A.5) is at most 5ϵ.

We are now in a position to complete the proof. The analysis so far shows that we can find a sequence of random variables $\hat{\xi}_n$ such that (A.5) converges to zero and such that $\hat{\xi}_n - \xi_n$ converges uniformly (in ω) to zero. An analogous argument shows that we can find $\hat{\eta}_n$ such that

$$\int \chi(\hat{\eta}_n - \underline{x}) \max\{0, -\frac{\partial b}{\partial x}(\hat{\eta}_n,\theta,0)\}\, dP_n(\omega,\theta) \tag{A.7}$$

converges to zero, and such that $\xi_n - \hat{\eta}_n$ converges uniformly to zero. Next, passing to a subsequence if necessary, we may assume that $P_n \to P_\infty$ for some P_∞. Also, by Skorohod's theorem, we can find a sequence of random variables (U_n,T_n,ζ_n) on an auxiliary probability space such that: (A.5) can be represented as

$$\bar{E}\left[\chi(\bar{x} - \hat{\xi}_n(U_n))\max\{0,\frac{\partial b}{\partial x}(\hat{\xi}_n(U_n),T_n,0)\}\right]; \tag{A.8}$$

an analogous representation holds for (A.7); and the expectation of the asymptotic conditional payoff can be represented as

$$\bar{E}(b(\xi_n(U_n),T_n,\zeta_n)). \tag{A.9}$$

Moreover $(U_n,T_n,\zeta_n) \to (U_\infty,T_\infty,0)$ a.s., where (U_∞,T_∞) has distribution P_∞. But (A.8) and its analogue imply that, with probability one, every limit point of $\xi_n(U_n)$ maximises

$b(\cdot,\theta,0)$. Hence (A.9) converges to $\bar{E}(b(X(T_\infty),T_\infty,0)) = \int b(X(\theta),\theta,0)dP_\infty(\omega,\theta)$. Finally, P_n is a measure on $\Theta \times Z^\infty \times \Theta$. Because Q_n is a conditional probability measure, the marginal of P_n over the second Θ is Q. Hence the marginal of P_∞ over this Θ is Q. Hence $\int b(X(\theta),\theta,0)dP_\infty(\omega,\theta) = \int b(X(\varphi),\varphi,0)dQ(\varphi)$. But this last integral is precisely the complete-information payoff when $R = \delta_0$. □

Proof of Theorem 4.3

Fix N and consider the strategy: play each ξ_n for $1 \le n \le N$ a total of T_N times, where T_N is as in the proof of Theorem 3.3; estimate $B(\xi_n,\theta)$ by averaging; and play the highest estimate $\xi_{i(N)}$ for ever. Then the payoff from this strategy satisfies:

$$(1-\delta)E\left[\sum_{t=1}^{\infty}\delta^{t-1}\pi_t\right] \ge (1-\delta^{NT_N})(-K) + \delta^{NT_N}\left[(1-N^{-2})E[M_N(\theta) - \tfrac{1}{N}] - KN^{-2}\right].$$

Call the latter quantity $\gamma_N(\delta)$. The limit infinum of the optimal payoff when δ tends to 1 is at least the limit infimum of $\gamma_N(\delta)$. Therefore, for any $N \ge 1$, the limit infinum of the optimal payoff is at least

$$(1-N^{-2})E[M_N(\theta) - \tfrac{1}{N}] - KN^{-2}$$

which implies that it is equal to $E[M(\theta)]$. □

REFERENCES

Alchian, A (1950): "Uncertainty, Evolution and Economic Theory", Journal of Political Economy 58, pp 211–221.

Alpern, S (1985): "Search for Point in Interval, with High–Low Feedback"; Mathematical Proceedings of the Cambridge Philosophical Society 97, pp 569–577.

——— & D J Snower (1987a): "Inventories as an Information–Gathering Device"; Theoretical Economics Discussion Paper (ST/ICERD), no 151, London School of Economics.

——— & ——— (1987b): "Production Decisions under Demand Uncertainty: The High-Low Search Approach"; Centre for Economic Policy Research Discussion Paper no 223, London.

——— & ——— (1988): "High–Low Search in Product and Labour Markets"; American Economic Review 78, pp 356–362.

Clower, R W (1959): "Some Theory of an Ignorant Monopoly", Economic Journal 69, pp 705-716.

Easley, D & N Kiefer (1988): "Controlling a Stochastic Process with Unknown Parameters"; Econometrica 56, no 5, pp 1045–1064.

Gihman, I I & A V Skohorohod (1979): Controlled Stochastic Processes; Springer–Verlag.

Grossman, S J, R E Kihlstrom & L J Mirman (1977): "A Bayesian Approach to the Production of Information and Learning by Doing"; Review of Economic Studies 44, pp 533–547.

Kiefer, Nicholas M (1987): "Optimal Collection of Information by Partially Informed Agents"; Economic Review, forthcoming.

Kihlstrom, R, L Mirman & A Postlewaite (1984): "Experimental Consumption and the 'Rothschild Effect'"; in Bayesian Models of Economic Theory, M Boyer and R E Kihlstrom (eds), Elsevier, Amsterdam.

Lazear, E P (1986): "Retail Pricing and Clearance Sales", American Economic Review 76, pp 14–32.

McLennan, A (1984): "Price Dispersion and Incomplete Learning in the Long Run", Journal of Economic Dynamics and Control 7, pp 331–347.

Parthasarathy, K R (1967): Probability Measures on Metric Spaces; Academic Press.

Reyniers, D (1989a): "A High–Low Search Algorithm for a Newsboy Problem with Delayed Information Feedback"; Operations Research, forthcoming.

——— (1989b): "Interactive High–Low Search: The Case of Lost Sales"; Journal of the Operational Research Society 40, no 8, pp 769–780.

Rob, Rafael (1988): "Learning and Capacity Expansion in a New Market under Uncertainty"; Working Paper, University of Pennsylvania.

Rothschild, M (1974): "A Two–Armed Bandit Theory of Market Pricing"; Journal of Economic Theory 9, pp 185–202.

Striebel, C (1975): "Optimal Control of Discrete–Time Stochastic Systems"; Lecture Notes in Economics and Mathematical Systems 110, Springer–Verlag.

Yosida, K (1968): Functional Analysis; Springer–Verlag.