Optimal Learning through Experimentation by Microeconomic Agents

R. Godfrey Keller

London School of Economics & Political Science

Thesis submitted for the degree of Doctor of Philosophy

Abstract

This thesis concerns itself with optimal learning through experimentation by microeconomic agents.

The first part presents a model of a search process for the best outcome of many multi-stage projects. The branching structure of the search environment is such that the pay-offs to various actions are correlated; nevertheless, it is shown that the optimal strategy is given by a simple reservation price rule. A simple model of R&D is provided as an example.

These general results are then applied in a model of job search and occupational choice in which jobs are grouped into occupations in a natural way. Before getting a job, the agent must first become qualified in the chosen occupation, at which point his general aptitude for jobs in this occupation is revealed. The search environment is such that the returns of jobs are correlated within a given occupation, but the optimal strategy is given by the above reservation value rule. One implication of this is that young inexperienced workers prefer to try riskier jobs/occupations first. Issues of job turnover and expected returns are addressed.

The second part studies optimal experimentation by a monopolist who faces an unknown demand curve subject to random changes, and who maximises profits over an infinite horizon in continuous time. Two qualitatively very different regimes emerge, determined by the discount rate and the intensities of demand curve switching, and the dependence of the optimal policy on these parameters is discontinuous. One regime is characterised by extreme experimentation and good tracking of the prevailing demand curve, the other by moderate experimentation and poor tracking. Moreover, in the latter regime the agent eventually becomes 'trapped' into taking actions in a strict subset of the feasible set.

Contents

List of Figures

Acknowledgements

My greatest debts, in economics, are to Patrick Bolton and John Hardman Moore – the latter for demonstrating how captivating the subject could be and for 'getting me started,' the former for showing how fascinating it continued to be and for 'keeping me going.' I thank them both for their encouragement, guidance, and infectious enthusiasm. Margaret Bray also deserves an early mention, not least for her MPhil course on Learning in Economics which set me off down the road which this thesis travels and which led to my asking Patrick to be my PhD supervisor.

Of course, I also owe an enormous amount to Alison Hole and Sven Rady, my co-authors on Chapters 1 and 3 respectively, and, via them, to their supervisors at LSE, John Sutton and Lucien Foldes. Working with each of Alison and Sven has been both a pleasure and an education – many, many thanks, and I hope a good advertisement for collaborative research.[1]

I met Alison in my early days in the LSE Financial Markets Group (FMG), whose reputation as a great working and social environment is well-known and well-deserved. (I am more than grateful to Reinout Koopmans for introducing me to the FMG, and to David Webb, its generous director.) When our joint work (which became Chapter 1) was nearing completion and I was contemplating a second strand (destined to be Chapter 3), she suggested that Sven may well be interested and might make a useful contribution – what an understatement! Moreover, both projects benefited from early airings at the informal forum provided by the FMG student seminar, run by Margaret Bray, whose efforts are particularly appreciated.

Kevin Roberts nobly deputised as my supervisor when Patrick was ill and I was in Edinburgh (thanks to John Moore), before Patrick suggested that dialogue would be easier if I were to join him at ECARE in Brussels, where Mathias Dewatripont was a welcoming host.

[1] When asked about the contribution of Morgenstern to their *magnum opus*, von Neumann reputedly replied: "I couldn't have done it without him." As I am the sole author of Chapter 2, signal-extraction techniques might help to answer the re-posing of that question in the context of this thesis.

Funding in the early days was from savings, so I must thank Michael Berman, who ran the computing company I used to work for: not only did his reference help me receive an offer from LSE, but also the salary I had been earning over the years enabled me to accept it. Subsequently, I received financial support, for which I am grateful, from a number of sources, principally from the ESRC through the Financial Markets Group at LSE and from the European Commission's Human Capital & Mobility Programme (Contract No. ERBCHBG-CT94-0595) through ECARE.

People whom I met and shared offices with at LSE and ECARE have become long-standing friends: Alison Hole, Sven Rady and Reinout Koopmans have already been mentioned; to these should be added Mike Burkhart and Vittoria Cerasi in London, Alain Desdoigts, Guido Friebel and Mike Raith in Brussels, and many others in a list from LSE, ECARE, and beyond, which continues to grow.

And then there is Denis Gromb ... When I inquired as to the whereabouts of my copy of Blanchard & Fisher (London? Brussels? Boston?) he promptly replied that he had deliberately lost it for my own good – I should stick with microeconomics – and he handed me his dog-eared LaTeXmanual in exchange. Thanks.

Finally, my parents and Tanya ... When they 'assisted' me through my first doctoral thesis 20 years ago I presume they didn't think that they'd have to do it again. As my father didn't quite make it to the end, this piece of work is dedicated to his memory. Warm thanks to them all, now, as then. (I suppose a third one is out of the question?)

London
1998

Introduction

In many areas of human activity, an agent has to choose from a number of actions, each with a cost and an uncertain reward. Some of these actions are highly likely to produce a short-term gain, while others, such as gathering information to eliminate some of the uncertainty, may result in only a long-term benefit. A firm engaging in R&D is unsure about the products which might emerge, and what the demand for those products might be. Even after product launch, consumer tastes are likely to change, and competitors enter the market. An individual may be unclear where his talents lie – should he look for another job, or maybe pursue a different career? These are examples of the sorts of question addressed in this thesis.

The classic multi-armed bandit problem is one formalisation of such a situation: in each period the agent pays a unit cost to pull one of a fixed number of arms, different arms having different and unknown pay-offs. When projects are equated with arms, there is no ambiguity about how to engage a project: with just one arm per project the only available action is to pull it; further, taking an action leaves the number of possible actions unchanged: with still just one arm per project the only available action is to pull it again. However, many decision environments are more complex.

Consequently, in Part I we introduce a model of a more general sequential search process in which, when an action is taken in one period, several new actions become available in the next period. The set of projects and the actions available within them depend on the previous choices of the agent.

Even the classic multi-armed bandit problem resisted any general solution until Gittins and his co-workers showed, in a very general setting, that if the arms are independent then the optimal strategy is given by an index policy, or reservation price rule. Calculating the indices, however, can be a formidable task.[1] Models in which the independence assumption is dropped have no simplifying result comparable to that of Gittins to help in determining the optimal strategy. Indeed, subsequent to

[1]Two notable applications in the economics literature of bandit problems with independent arms are Weitzman [1979] and Roberts & Weitzman [1981], in which the examples focus on cases where the reservation price is not so difficult to calculate.

the paper by Rothschild [1974] which introduced bandit problems into the economics literature, work on similar pricing problems has abandoned the bandit terminology altogether, and the usage of the term 'bandit' *now* appears to be reserved for cases where the different arms are independent.

Part I introduces a general sequential search process in which the possible actions belong to branching projects. This process generalises a standard multi-armed bandit in a number of significant ways: an action can reveal information about more than one reward; the pay-offs to various actions are correlated; and there is a natural way to talk about the diversity of rewards. We give a simple characterisation of when the independence assumption can be relaxed, but with the problem retaining the analytical convenience of the optimal strategy being determined by an index policy or reservation price rule.

A branching project is special case of a multi-action project, a project in which there may be several alternative actions which the agent can take at any one time, and where this set of available actions depends on the agent's previous choices. A Gittins index can be attached to a multi-action project in much the same way as to a single-action project, but generally, when the index policy is optimal, it does not specify the optimal action, only the project to which the action belongs. In the special case where the multi-action projects are branching projects we give a condition under which the Gittins index policy picks out not only the optimal project to engage but also the optimal action within that project. (Essentially, this condition is that taking one action gives no information about actions which do not emanate from it.)

The optimality of the Gittins index policy for a class of branching projects considerably reduces the problem of characterising the optimal search strategy, and in Chapter 1 we use a simple model of R&D in order to demonstrate the usefulness of our result, deriving the optimal strategy in a generalised way and discussing some of its features.

In Chapter 2, we particularise the general model to one of job search and occupational choice, and extend the detailed analysis to incorporate discounting and 'earn-as-you-go'. Here, a job is treated as an *experience* good, that is, the agent finds out about the value to him of a particular job only after being hired.[2] However, before looking for a job in a particular occupation, the agent must first become

[2]Contrast this with models of Job Matching that treat jobs as *inspection* goods, where the value of the match is revealed prior to the match being proposed, in, for example, Diamond [1982] and Pissarides [1990].

Other matching models of job search which treat jobs as experience goods are to be found in Jovanovic [1979], Miller [1984], and Felli & Harris [1996].

qualified for that sector/ profession/ trade, which is also an *experience* good – the training stage reveals the agent's general aptitude for jobs in the given occupation. The precedence constraints and costly information revelation at each stage make the underlying model appropriate, as does the correlation of the returns to the agent within an occupation.

Accordingly, we can determine the optimal behaviour of the agent, i.e. which occupation to acquire skills in first, when to change jobs, whether or not to retrain. Further, we can address questions such as: How many jobs will the typical agent try before settling down? What return stream can the agent expect to end up with? What is the probability of the agent finding the most suitable job in a given occupation?

It will transpire that new entrants to the job market will rationally launch themselves into professions in which the returns are more risky, where the average return is on the low side and where turnover is high, whereas those with more experience will have found themselves jobs that are suitable enough so that it is not worth their while to look elsewhere. So, we offer an alternative explanation of this behaviour: it need not be that the young are impetuous, have a presumption of success, or that they are overoptimistic and have unrealistic expectations.

Part II consists of a single chapter. The problem which the agent faces is still the one of trading-off the short-term opportunity cost of his action against the long-term informational benefits, but it differs from the problem analysed in Part I in a number of ways. First, the action space is continuous and it is not a bandit problem. Secondly, the costs are implicit and noisy, not explicit and known; also, the informational benefits are noisy. Finally, the underlying environment is not fixed, in that we allow it to change over time.

In Chapter 3, we consider an economic agent whose per-period rewards depend on an unobservable and randomly changing state. Owing to noise, the reward observed after taking an action provides only an imperfect signal of the prevailing state, and the agent can improve the information content of this signal by experimenting, that is, by deviating from the myopically optimal action that just maximises current pay-off. The long-term benefits of experimentation are more informed decisions in the future; its short-term opportunity cost is the pay-off forgone in the current period.

We are interested in a number of issues. How does the agent's optimal action differ from what is myopically optimal, and is this difference large or small? How well does the agent track the prevailing state? We address these questions in a setting where the agent can finely control the information content of the signals he

receives, over a range from zero to some natural upper bound.

Our main result is the identification of two qualitatively very different experimentation regimes. One regime is characterised by large deviations from myopic behaviour, guaranteeing that the signals observed by the agent always contain at least a certain amount of information. This allows him to track the state well, in the sense that his beliefs can come arbitrarily close to the truth. The other regime is characterised by small deviations from myopic behaviour, resulting in signals whose information content can become arbitrarily small. In this regime, the prevailing state is tracked poorly: the agent eventually becomes 'trapped' in a strict subset of actions such that, in one of the states, beliefs always stay bounded away from the truth.

Specifically, the agent in the model of Chapter 3 is a monopolist facing an unknown and changing demand function and maximising expected profits over an infinite horizon. There are two possible states, each characterised by a linear demand curve, and the transitions between these states are governed by a Markov process. The monopolist knows the slope and intercept of each demand curve and the transition probabilities, but he does not know which demand curve he faces at any given time. At each instant, he chooses from a given interval of feasible quantities, and observes a price which is the 'true' price (derived from the prevailing demand curve) plus noise. Given this noisy signal of the underlying state, the monopolist updates his belief in a Bayesian fashion.

The monopolist can increase the information content of the price signal by moving away from the confounding quantity, that is, the quantity at which the two demand curves intersect; setting the confounding quantity itself leads to a completely uninformative signal. Focusing on the most interesting case, we assume that the confounding quantity lies between the quantities which are myopically optimal in each of the two states. This implies that there is a unique belief – the confounding belief – at which the confounding quantity would be chosen by a myopic agent. The two experimentation regimes are distinguished by the optimal behaviour near this belief.

For a given level of noise, when the discount rate and the intensity of state switching are both low, then experimentation is *extreme*: for beliefs in an interval encompassing the confounding belief, the optimal action is to choose a quantity at the boundary of the feasible set, and the optimal quantity (as a function of the belief) exhibits a jump from one boundary to the other. In this regime, the agent's belief tracks the true state well.

When, for the same level of noise, either the discount rate or the switching

intensity is high, then experimentation is *moderate*: the monopolist chooses the confounding quantity at the confounding belief, and quantities relatively close to the myopic ones everywhere else. In this regime, the monopolist eventually becomes trapped into choosing quantities on just one side of the confounding quantity. Then, the continually changing state entails his belief sometimes being on the 'wrong' side of the confounding belief, in which case it can never get closer to the true state than the confounding belief.

Thus, in this scenario, optimal behaviour depends *qualitatively* on the switching intensities, and a small change in the likelihood of switches can trigger a *discontinuous* change in the optimal policy. Thus, a small increase in the variability of the environment may not just lead to a moderate reduction in information gathering activities – in fact it could provoke a near cessation of these activities, with drastic consequences for the process of information aggregation.

We build upon several strands of the literature on optimal Bayesian learning. First, a number of authors have identified situations where it is optimal to experiment, and have characterised the agent's strategy as a function of his beliefs.[3] These papers do not consider confounding actions, so the different experimentation regimes described here do not arise.

Working in an infinite-horizon setting where the unknown reward function is fixed over time, other authors have focused on the agent's limiting behaviour.[4] A common result of these papers is that the agent's beliefs and actions converge. In the limit, the agent learns everything that is *worth* knowing, so experimentation ceases and no further information is gathered. If there is a confounding action and the agent is impatient, however, beliefs need not converge to a one-point distribution at the true reward function, i.e. learning can remain incomplete. Our moderate experimentation trap extends this incomplete learning result to a changing environment.

Allowing the reward function to change randomly adds more realism in that new data continues to be pertinent, so beliefs continue to evolve, and the agent is not doomed to take the same action for evermore. Moreover, the prior with which the agent starts becomes irrelevant in the long run. Here, we follow Kiefer (1989b), Bala and Kiefer (1990), Balvers and Cosimano (1990, 1993, 1994), Rustichini and Wolinsky (1995) and Nyarko and Olson (1996). However, these authors have either

[3]Examples include Prescott (1972), Grossman, Kihlstrom and Mirman (1977) and, more recently, Bertocchi and Spagat (1993), Leach and Madhavan (1993), Mirman, Samuelson and Urbano (1993) and Trefler (1993).

[4]The first such model in the economics literature is due to Rothschild (1974), and has subsequently been extended in a number of different directions; see, for example, McLennan (1984), Easley and Kiefer (1988), Kiefer (1989a), and Aghion, Bolton, Harris and Jullien (1991).

focused on different aspects of the problem, or used frameworks that lent themselves to only limited analytical investigation. The two papers closest to ours are Kiefer (1989b)[5] and Balvers and Cosimano (1990),[6] both studying a monopolist learning about changing linear demand curves, but see also Rustichini and Wolinsky (1995).[7]

We depart from the above papers by formulating the problem in continuous time. The advantage of this approach is that it allows us to derive sharp analytical results. We are able to establish key properties of the value function and the optimal policy; we obtain some analytical comparative statics results; and it is straightforward to characterise the sample path properties of beliefs and optimal actions in each of the two experimentation regimes.

Continuous-time models in the economics literature on Bayesian learning have been pioneered by Smith (1992) and Bolton and Harris (1993), and pursued by Felli and Harris (1996). Building on a bandit structure as in Karatzas (1984) and Berry and Fristedt (1985), these authors examine multi-agent learning problems with a fixed distribution of rewards.[8] We follow these three papers with our specification of Brownian noise and the reliance on the filtering techniques from Liptser and Shiryayev (1977). There are two major differences, however: the problem we study is not of the bandit type, and we allow for a changing environment.

[5] In a framework with two possible demand curves, Kiefer calculates the value function numerically, illustrates two types of optimal policy (one continuous, one with a jump) and simulates the corresponding sample paths of beliefs and actions.

[6] In Balvers and Cosimano's framework, both intercept and slope of the unknown demand curve are given by stochastic processes, so there is in fact a continuum of possible demand curves. The added complexity makes it very hard to obtain analytical results, and moreover, the absence of a confounding action means that their result of sluggish price adjustments has no direct comparison with our main findings.

[7] Rustichini and Wolinsky use a two-armed bandit framework to study monopoly pricing when the buyers' reservation value changes randomly. Their focus is on non-negligible pricing errors even when the frequency of change is negligible. For certain parameter combinations, learning will cease completely even though the state keeps changing. This can be seen as the analogue in a discrete action space of our moderate experimentation trap.

[8] Smith considers agents that enter sequentially and learn by observing a 'snapshot' of the actions taken by previous generations. He shows that the incomplete learning result going back to Rothschild (1974) is not robust to this form of market learning.

Whereas Smith's model precludes strategic behaviour (agents do not observe each other once they have entered), Bolton and Harris focus on the informational externality arising when several agents experiment simultaneously and observe each other's actions and outcomes.

Felli and Harris use a variant of the continuous-time bandit framework to study equilibrium wage dynamics in a setting where two firms and a worker learn about the worker's aptitude to perform firm-specific tasks.

Part I

Exploring Branching Structures

Chapter 1

A Bandit Problem with Correlated Pay-offs

Introduction

In many areas of human activity, an agent has to choose from a number of actions, each with a cost and an uncertain reward. Some of these actions are highly likely to produce a short-term gain, while others, such as gathering information to eliminate some of the uncertainty, may result in only a long-term benefit. The classic multi-armed bandit problem is a formalisation of such a situation: in each period the agent pays a unit cost to pull one of a fixed number of arms, different arms having different, unknown, and possibly interdependent pay-off probabilities; the agent's problem is to maximise the expected discounted sum of pay-offs.

In bandit problems currently in the economics literature, projects are equated with arms. There is no ambiguity about how to engage a project: with just one arm per project the only available action is to pull it. Further, taking an action leaves the number of possible actions unchanged: with still just one arm per project the only available action is to pull it again. However, many decision environments are more complex. Here we introduce a model of a more general sequential search process in which, when an action is taken in one period, several new actions become available in the next period. The set of projects and the actions available within them depend on the previous choices of the agent.

Even the classic multi-armed bandit problem resisted any general solution until Gittins and his co-workers showed, in a very general setting, that if the arms are independent (that is, pulling one arm is uninformative about other arms) then the

optimal strategy is given by an index policy.[1] To each arm attach an index (known variously as a reservation price, dynamic allocation index or Gittins index) which depends on the current state of only that arm; the strategy is to pick the arm which currently has the highest index. Calculating the indices, however, can be a formidable task. In the economics literature, two notable applications of bandit problems with independent arms are Weitzman [1979][2] and Roberts & Weitzman [1981],[3] in which the examples focus on cases where the reservation price is not so difficult to calculate.

Models in which the independence assumption is dropped have no simplifying result comparable to that of Gittins to help in determining the optimal strategy. Nevertheless, the paper by Rothschild [1974][4] which introduced bandit problems into the economics literature centres on an example of such a model, and he derives strong results on how much a monopolist learns about a stochastic demand function. Subsequent work on similar pricing problems[5] has abandoned the bandit terminology altogether, and indeed the usage of the term bandit *now* appears to be reserved for cases where the different arms are independent.

In this chapter, we introduce a general sequential search process in which the possible actions belong to branching projects. This process generalises a standard multi-armed bandit in a number of significant ways: an action can reveal information about more than one reward; the pay-offs to various actions are correlated; and there is a natural way to talk about the diversity of rewards. We give a simple characterisation of when the independence assumption can be relaxed, but with the problem retaining the analytical convenience of the optimal strategy being determined by an index policy or reservation price rule.[6]

---

[1] See the references to papers by Gittins, Glazebrook, Jones and Nash here and in Whittle [1980].

[2] Weitzman considers a problem where there are several substitutable single-stage projects, which can be sampled sequentially. When the agent decides to stop searching, only one option is selected, namely the one with the maximum sampled reward.

[3] Roberts & Weitzman look at an application to R&D in which there is a single multi-stage project. Costs are additive (pay-as-you-go), benefits are received only at the end, and the choice facing the agent at each stage is whether to pay to resolve more of the uncertainty *and* bring the project closer to completion, or to abandon the project.

[4] In this well-known paper, Rothschild models the pricing decision of a monopolist facing an unknown stochastic demand as a two-armed bandit problem. No assumption is made that the parameters governing demand at the two prices are independently drawn and Rothschild does not derive the optimal strategies. The main result is that optimal experimentation may not result in adequate learning, that is, there is a positive probability that after some finite period the agent will settle for the inferior arm for ever more.

[5] See, for example, Aghion, Bolton *et al.* [1991, §6], and the references in their introduction.

[6] Gittins [1989] uses the example of job scheduling with precedence constraints to motivate an abstract model which is a finite horizon version of that which we present in this chapter, but without the information revelation aspects or the reward correlation which we have here. Our

A branching project is special case of a multi-action project, a project in which there may be several alternative actions which the agent can take at any one time, and where this set of available actions depends on the agent's previous choices. A Gittins index can be attached to a multi-action project in much the same way as to a single-action project. In an extension of his proof of the original result (see Gittins & Jones [1974] and Gittins [1979]), Whittle [1980] gives a condition under which the Gittins index policy is optimal for multi-action projects; note that it does not specify the optimal action, only the project to which the action belongs. In the special case where the multi-action projects are branching projects we give a condition under which the Gittins index policy picks out not only the optimal project to engage but also the optimal action within that project. Essentially, this condition is that taking one action gives no information about actions which do not emanate from it.

The optimality of the Gittins index policy for a class of branching projects considerably reduces the problem of characterising the optimal search strategy. We use a simple model of R&D in order to demonstrate the usefulness of our result, deriving the optimal strategy in a generalised way and discussing some of its features.

In the next section, we present the example of R&D in order to illustrate some of the features which branching projects possess and introduce some notation. Then in Section 2 we give a formal description of the general model, and the central theoretical result as a corollary of Whittle's theorem. In Section 3, we apply it to the model of R&D and provide some results and examples. We conclude with a discussion and some remarks. Proofs of the main technical results are to be found in the appendices.

1 Example: a simple model of R&D

A simple branching project is represented in Figure 1.1 by a tree, with node 1 as its root and nodes 4 through 7 as its terminal nodes. When there is an arc from node p down to node q we say that node p is a parent of node q and that node q is a child of node p. The terms ancestor and descendant have the obvious meanings.

The nodes correspond to possible actions, a subset of which are available in any given period. There are two sorts of possible action: one is to pay a cost c_n to explore node n and then continue; the other is to collect a prize whose value is y_n and which is located at an *explored* terminal node n, and stop. The actions which are available in any period depend on previous actions and can be summarised using

proof of the optimality of the Gittins index policy in this set-up was arrived at independently and adopts what we believe is a self-contained and accessible approach.

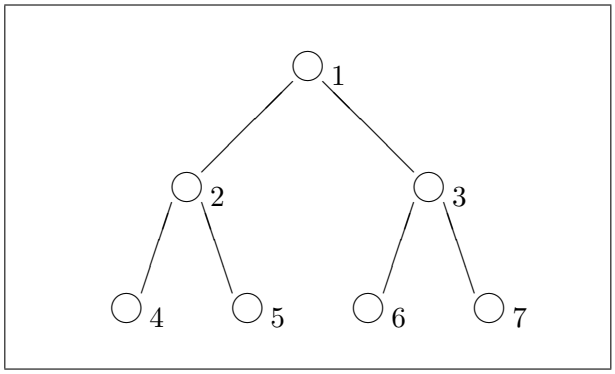


Figure 1.1: No nodes explored

the tree. We assume that initially no node has been explored, and now in any period the agent can (a) explore any node that has not yet been explored, provided that either it is the root or its parent has been explored, and then continue, or (b) collect the prize at a terminal node that has been explored and stop.

We shall often consider there being an additional fall-back option available in any period, and if it is chosen the agent collects a prize of value m and stops. For example, suppose that the situation is as illustrated in Figure 1.2, in which filled nodes have been explored and empty ones have not, and there is a fall-back. The available actions are: explore node 3, explore node 4, take y_5, or take the fall-back m.

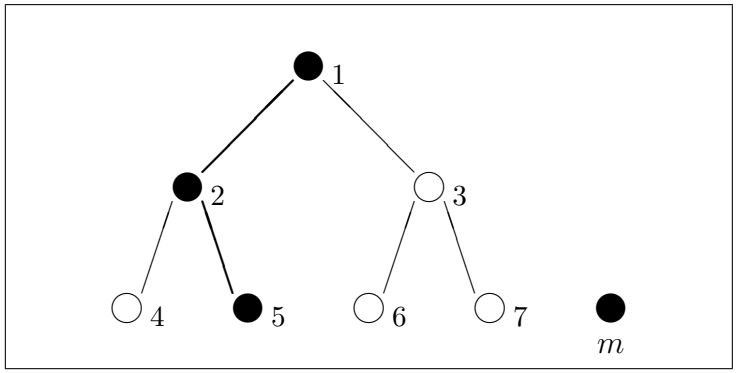


Figure 1.2: Some nodes explored, & a fall-back

In an R&D setting, node 1 might represent a feasibility study, and nodes 2 and 3 would represent two different avenues of basic research, each of which leads to two development opportunities. One would then think of nodes 4 through 7 as representing substitutable technologies to produce a product. To take the fall-back option is to use the existing technology, and abandon R&D. Note that 'production'

is also a terminating action – it corresponds to stopping R&D and commercially exploiting the know-how that has been gained.

Exploring a node not only imposes costs on the agent and affects which actions are available in future periods, but also reveals information about the prizes at all its descendent terminal nodes: when the agent explores node n she receives a random signal z_n, which is independent for each node. The value of the prize at a terminal node is the sum of the signals at that node and its ancestors, so, for example, $y_5 = z_1 + z_2 + z_5$. (Because the signals contribute additively to the prize, we sometimes refer to them as *increments*.) The implication of this for the model of R&D is that each piece of basic research is informative only about products which embody that research, and that developing one product is uninformative about the value of other products. This means that, whenever the agent updates the expected value of any product, she uses only what has been learnt at its explored ancestors.

The agent's problem is to choose a strategy which maximises the expected value of the prize that she collects when she stops, net of the expected costs from exploring nodes before she collects the prize.

Note that the way in which actions become available leads to a natural measure of the diversity of prizes: those with a common parent are closer than those with only a common grandparent. Moreover, as a result of the specification of the prizes themselves, the values of closer prizes are more correlated.[7] Two features of this example worth stressing are that in any period each available action can be considered as the root of its own *separate* and *independent* sub-tree. Reconsider the situation illustrated in Figure 1.2.

We can in fact represent the agent's choice as between the projects shown in Figure 1.3 in which each project now contains only one available action: explore an unexplored root and continue, or collect a prize and stop. This representation is legitimate because all the ancestors of currently available actions have been explored, and we can use the state of each project to effectively summarise the signals received at the ancestors of its root (and at the root itself, if it has in fact been explored). Further, these separate projects are independent: nothing that is subsequently learnt in one project reveals anything about the prizes available elsewhere, an inherited property that follows from the fact that the signals received at one node are informative about the prizes only at terminal nodes which descend from it.

[7] At the start, before the agent has received any signals, the values of all prizes are correlated random variables: they all depend on the realisation of z_1. The values y_4 and y_5 are closely correlated because $\mathrm{Cov}(y_4, y_5) = \mathrm{Var}(z_1) + \mathrm{Var}(z_2)$, and even when z_1 has become known they are still correlated. Contrast this with y_5 and y_6: $\mathrm{Cov}(y_5, y_6) = \mathrm{Var}(z_1)$, and once z_1 has become known they are uncorrelated.

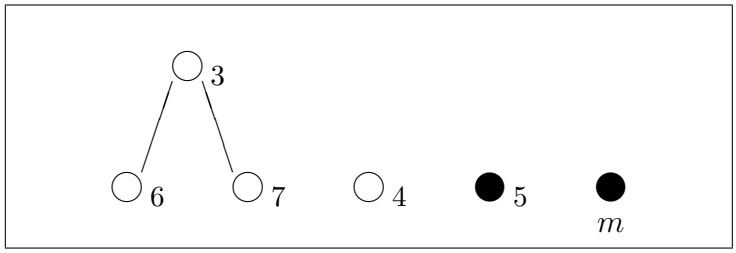


Figure 1.3: Separate and independent sub-trees

With regard to an index policy, were the agent to be in the above situation and treat the whole tree as a single project as in Figure 1.2, then a rule which selected the project with the highest index would simply tell the agent whether to proceed with the project or to take the fall-back. However, if she views the process with the perspective provided by Figure 1.3, and applies the rule to these separate projects, the strategy is completely characterised because just one action is picked out. Further, as we shall show, the fact that these separate projects are independent ensures that the Gittins index policy is optimal.

2 The General Model – Optimality of the Gittins Index Policy

In this section we develop more formally the central model of the chapter: a sequential decision process in which the alternative projects are branching projects. We introduce our definition of a branching project and state our result (Claim 2.1) that if the agent is choosing an action from among a set of independent branching projects then the optimal *action* in each period is given by the Gittins index policy. This is shown to be a corollary of a more general result on stationary Markov decision processes (Theorem 2.1) which gives the conditions under which the Gittins index policy picks out the optimal multi-action *project* to engage in each period.

2.1 Branching Projects

Borrowing some notation from graph theory, we represent a branching project by an out-tree,[8] in which the number of nodes may be infinite, but such that the out-degree

[8]Consider a directed graph, which is a set of nodes and a set of arcs, each arc being an ordered pair of nodes. An out-tree is a connected directed graph with a single root, no circuits and in which each node has no more than one parent.

of any node is finite, i.e. the tree can have infinite depth but only finite branching. The nodes are the actions within the project, and the arcs represent precedence constraints: an action (other than the root) can be taken only if its parent action has previously been taken. An action is available if it has not previously been taken, and either it is the root or it is the child of an action which *has* previously been taken.

We shall consider a family of branching projects, and in each discrete period, a risk neutral agent chooses one project and an available action within it. We first note that the set of alternative projects need not be the same in every period.

Lemma 2.1 *Consider a family of N branching projects. In every period, there is a partition of the actions which have not yet been taken into a set of branching projects in which only the root action is available.*

PROOF: That such a partition exists initially is clearly the case, so assume that such a partition exists at time t. If the agent engages project k by taking its root action then each of the children of that root is an action available at time $t+1$ and is the root action of a distinct sub-tree, none of whose actions have been taken. Also, each of the projects which was not engaged at time t is still a branching project in which only the root action is available. Hence such a partition exists at time $t+1$, and the lemma is proved by induction. ■

When project k is engaged by taking action u, the agent receives a reward and observes a signal, the signal affecting what the agent knows about the rewards associated with actions that may be available in later periods. The state of the project, denoted by x_k, is a sufficient statistic for the observational history. It summarises what has been learnt from past signals about future rewards, availability of actions, etc. and both the reward, $R_k(x_k, u)$, and the signal, $S_k(x_k, u)$, depend on the current state and the action taken. The new state of a project depends only on the old state and the action taken, both directly and indirectly via the signal. If signals are informative only about the rewards at descendent actions,[9] then the branching projects are *independent*, i.e. the state of unengaged projects remains unchanged.

Lemma 2.2 *Consider a family of N independent branching projects. If, after each period, the actions which have not yet been taken are repartitioned as in Lemma 1, then the branching (sub-)projects remain independent.*

[9]Let u be the action taken, and u' be any action which is not a descendant of u. The agent's expectation of the reward to be obtained from taking action u' is unchanged by the signal received from taking action u.

PROOF: Consider the partition at time t, and observe that no action in one project is a descendant of the root action of another. So when taking an action is uninformative about actions which do not descend from it, engaging any project by taking its root action is uninformative about other projects, and the lemma follows. ■

The importance of the above two lemmas lies in the fact that when an action in a project is taken, the state of the project changes but thereafter the action does not affect the agent's choices or pay-offs, so that in each period we need consider only those actions which have not yet been taken. The lemmas then imply that, if we start with independent branching projects, in each period we can view the agent as choosing between actions in a family of branching (sub-)projects which are still independent and in each of which there is just one action available, namely the root action. This is at the heart of Claim 2.1 below.

The agent's problem

Rewards are additive and discounted in time by a factor β, so the agent's problem is to choose a strategy to maximise the expected discounted sum of rewards from this process, whose state at time t is written as $x(t) = \langle x_1(t), x_2(t), \ldots, x_N(t)\rangle$. The maximal expected reward over feasible policies π, denoted by the value function $F(x)$, is given by:

$$F(x(0)) = \sup_{\pi} \mathrm{E}_{\pi}\Big[\sum_{0}^{\infty} \beta^t R(x(t), u(t)) \;\; | \;\; x(0)\Big],$$

where $R(x, u)$ is the immediate reward realised when action u is taken in state x. When the rewards are uniformly bounded, standard assumptions from dynamic programming are sufficient to establish that the value function is the unique bounded solution to the associated dynamic programming equation and that an optimal policy exists.[10]

2.2 Gittins Index Policy

Following the approach of Gittins and his co-workers, it can be shown that, under certain conditions, all optimal policies are contained in a simple class of policies, and the optimal action is that recommended by the Gittins index.

[10]Given that the rewards are additive, discounted in time by a factor β, and are uniformly bounded, the assumption that the agent is facing a stationary Markov process, for example, is sufficient.

Suppose that we can attach an index to any project k, that is a value $m_k(x_k)$ which is a function only of the project and its current state. When the agent selects the project with the currently highest index, she is said to be following an index policy. The specific index we shall look at is the Gittins index, whose definition makes use of a fall-back option. When there is a fall-back option m, then the agent has a stopping problem in which in each period (given that the fall-back option m has not yet been taken) the agent can either take the fall-back and stop, or continue the project for another period (the option of taking the fall-back remaining open in subsequent periods). The smallest value of m which makes the agent indifferent between stopping and continuation is the *Gittins index* of the project.

Denote the value function for the modified problem consisting of a fall-back M together with N projects by $\Phi(M, x)$. Since the rewards are bounded, we see that $\Phi(M, x) = M$ when M is large, and that $\Phi(M, x) = F(x)$ when $-M$ is large, and so the Gittins index is well-defined. The usefulness of this index is shown in the following result.

Optimality of the index policy for branching projects

Claim 2.1 *Consider a family of N independent branching projects in which the rewards are uniformly bounded.*

Then the Gittins index policy selects not only the best project to engage but also the optimal action within that project.

PROOF: Using Lemmas 2.1 and 2.2, after each period we can repartition the actions which have not yet been taken into independent sub-projects in each of which just the root action is available. The claim then follows as a corollary of the more general result for super-processes which we present in the next sub-section, because the two sufficient conditions for the theorem hold. Essentially these are: (a) the state of unengaged projects remains unchanged (because signals are informative only about descendent actions); and (b) the optimal action within the engaged project is independent of the size of the fall-back (because repartitioning after each period ensures that there is only one action available in each sub-project). The theorem then tells us that the project to which the optimal action belongs is the one with the highest Gittins index, and so the optimal action is the root action of the sub-project picked out by the Gittins index policy. ■

The above proof highlights the dual role of repartitioning actions into projects with only root actions available: it provides a key condition for the theorem, and it allows us to move immediately from 'best project' to 'optimal action'.

2.3 Bandit Super-processes

The proof of the above result relies on a theorem for super-processes which we present here.[11]

A super-process[12] is defined by the following collection:

(1) a set of projects, indexed by $k = 1, \ldots, N$;

(2.1) a state space, with generic element denoted by x;

(2.2) a set of available actions for each project when in state x, denoted by $U_k(x)$;

(2.3) a bounded real-valued reward function $R_k(x, u)$ which describes the instantaneous reward from taking action u in project k when in state x;

(2.4) a state transition rule giving the probability of next period's state, conditioned on this period's state, the action taken & the project it is in;

(3) a discount factor β.

The agent discounts the future by a factor β and aims to maximise the expected discounted sum of rewards from this process.

It is a *bandit* super-process when the state transition rule refers to each project rather than the process as a whole, and also when the action set and the reward are functions not of the *process* state but of the *project* state. (So, items (2.1) through (2.4) above would be for each project, and x should be replaced by x_k.)

Thus, given a bandit super-process, if project k is engaged in period t by choosing action $u \in U_k(x_k(t))$, the agent receives a reward of $R_k(x_k(t), u)$; states of unengaged projects do not change and the state of the engaged project changes by a Markov transition rule: if $j \neq k$ then $x_j(t+1) = x_j(t)$, and the value of $x_k(t+1)$ is conditioned only by $x_k(t)$, u & k.

We assume that the Markov process is *stationary* or time-homogeneous, i.e. the available action set, the reward, the state transition rule and the discount factor do not depend *explicitly* on time. (To give this some force, we do not allow time to be incorporated into the state.)

When the agent is maximising the expected reward from a super-process she must choose both which project to engage and which action to choose within that project. The theorem below shows that the Gittins index policy is optimal if two conditions are met: (a) projects are independent (i.e. it is a bandit super-process); (b) when there is a fall-back available, the optimal action within the engaged project is independent of the size of the fall-back.

[11] For a fuller treatment, see the appendices and the references cited there.

[12] The terminology is due to Gittins [1979], though the notion is due to Nash [1973]. However, Glazebrook [1982] uses 'super-process' to mean a multi-action project and so discusses a family of alternative super-processes.

Theorem 2.1 (Whittle) *Consider a super-process consisting of N alternative multi-action projects. Assume:*
(a) the projects are independent, i.e. the states of unengaged projects do not change;
(b) when there is a fall-back option available, the optimal action within the engaged project is independent of the size of the fall-back.

Then the Gittins index policy is optimal, in that it selects the best project to engage.

Moreover, writing $\phi_k(m, x_k)$ as the analogue of $\Phi(M, x)$ when only project k is available, the following identity holds:

$$\Phi(M, x) = B - \int_M^B \prod_k \frac{\partial \phi_k(m, x_k)}{\partial m}\, dm$$

where B is the bound on the reward functions.

PROOF: The proof is outlined in the appendices. Appendix A gives the proof for simple bandit processes (for which the second condition is vacuous), and Appendix B generalises it to bandit super-processes for which the second condition is crucial. The approach is essentially due to Whittle [1982] and the proof elaborates on that in Whittle [1980]. ■

It should now be clear from the definitions that a branching project is a super-process, and that a family of *independent* branching projects constitutes a *bandit* super-process, so the first condition for the theorem is met. Moreover, the lemmas show that it is legitimate to reorganise the available choices in a convenient way, so that not only is the second condition for the theorem met, but also the result is strengthened from the Gittins index selecting the best project in a general bandit super-process to it picking out the optimal action from a family of independent branching projects.

Notes

Given that a branching project can have only finite branching (although it can have infinite depth), after any finite number of periods there will be only a finite number of actions available. Thus a branching project here is *not* an *infinite-armed* bandit (as in, for example, Banks & Sundaram [1992]).

Also, the number of available actions does not increase spontaneously (in a Poisson stream, for example), but only after a deliberate action by the agent. Thus a branching project is *not* an *arm-acquiring bandit* (as in Whittle [1981, 1982]), and it may be more convenient *not* to think of a branching project as an "open" process (Whittle's terminology) even though the number of available (sub-)projects increases with time.

Further, irrespective of independence, we make the traditional assumption that a project is static in its passive phase, i.e. unengaged projects do not deteriorate nor improve, for example. This means that branching projects are *not restless bandits* (in the sense of Whittle [1988]).

2.4 Discussion

The index result reduces the original problem significantly: the index is calculated without reference to any other outside option or project, and the optimal action emerges from a comparison of the indices $m_k(x_k)$ attached to the various projects; further, the index of any unengaged project does not change, and so need not be recalculated. We should stress that the index is used to determine which project to engage next when the other projects will still be available in the next period. *It is not the expected value of the project.* A brief example will illustrate this point.

Consider two projects A and B. You must decide which project to engage first, and then whether you want to stop, or to engage the other project and take the larger pay-off. The cost of project A is 20 and it results in a pay-off of either 200 or zero, each outcome being equally likely. The cost of project B is 10 and it results in a pay-off of 170 or 130, again with each outcome being equally likely. So, the net expected value of project A is 80, and that of project B is 140. However, the Gittins indices for the projects are 160 and 150 respectively, so it is optimal to engage project A first, and only then engage project B if the low outcome prevails.[13]

It is to the calculation of the indices, or reservation prices, that we turn in the next section, after a few remarks on processes consisting of projects with variable length project stages, and on finite versus infinite horizon problems with discounting.

Variable length project stages

If projects have stages whose length can vary, we assume that when the agent engages a project she is committed to it for a possibly random number of periods, that number being dependent on the current state of the project but not on the actual period in which the stage was begun. As is indicated in the appendices, the proof of the optimality of the Gittins index policy continues to hold.

[13]This also demonstrates the principle that you should engage the riskier project first – the down-side is unimportant because you will never end up taking the low outcome from project A. This is shown more formally in Result 3.1 of the next section.

Finite versus infinite horizon, and discounting

There are two ways of looking at the fall-back m. The first is: in any period, *either* select from the available projects, *or* settle for a once-and-for-all lump sum pay-off of m and abandon selection for ever. The second is: in any period, *either* select from the available projects, *or* take a fixed reward of $(1-\beta)m$ this period and continue selection next period. In the latter case, if it is optimal to take the fixed reward of $(1-\beta)m$ this period, the agent learns nothing about the other projects, and so it is optimal to take the fixed reward of $(1-\beta)m$ in all subsequent periods, and the total discounted reward from this period forward is just m. Thus, in the infinite horizon case with discounting, the two views are equivalent.

Similarly, in the case when some projects have a terminating action,[14] if the agent selects such a project which is in a terminal state, this can be viewed as either settling for the associated lump sum reward, say y, and abandoning selection for ever, or as taking a fixed reward of $(1-\beta)y$ now (with the state of *all* projects remaining unchanged) and continuing selection next period. If we take the former view, this may seem to imply that the selection of a project which is in a terminal state affects the state of other projects because they are no longer available. However, if we redefine the fall-back as the maximum of m and y whenever a project reaches a terminal state with an associated lump sum reward of y, then once more the choice is between selecting from the available projects which have not yet reached a terminal state and taking the fall-back.

In the finite horizon case when all projects have terminating actions and there is no discounting, we are forced to take the former view (i.e. to take the fall-back is to settle for a lump sum pay-off of m and abandon selection for ever) and the last remark (i.e. redefinition of the fall-back whenever a project reaches a terminal state) applies.

3 Reservation Prices – Results and Examples

This section returns to the example of the project that was introduced in Section 1 and employs the interpretation of it as a model of R&D. Using the results just derived, we characterise the optimal strategy, and then discuss some implications of this strategy. Figure 1.4 illustrates the project. It differs from Figure 1.1 in that, to be more consistent with the exposition of Section 2, the new figure also shows the

[14] This corresponds to the notion of *stoppable* super-processes in Glazebrook [1982]. The simple model of R&D presented in Section 1 is an example of such a process.

actions of costless production (nodes $4'$ through $7'$). Also, although the figure only ever shows two branches, we may wish to assume that in the project itself there are more, and denote the number of branches by γ.

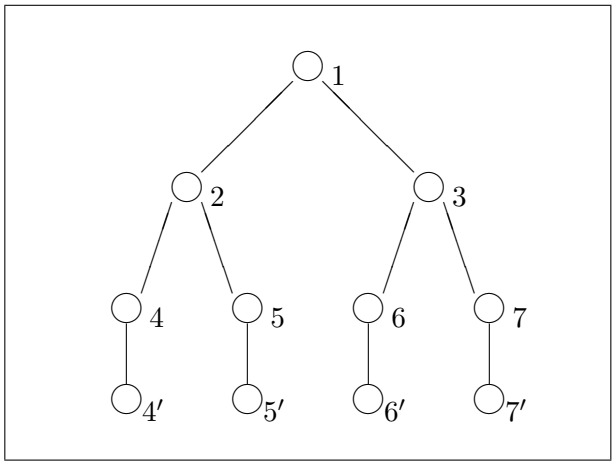


Figure 1.4: The project

3.1 Characterising the Optimal Strategy – Gittins Indices

The project is clearly an independent branching project, in which the only action initially available is the root, and the out-tree which describes the structure is the set of arcs illustrated in Figure 1.4. As noted after Theorem 2.1 in the previous section, this means that a Gittins index policy selects the optimal action, and so to characterise the optimal strategy we need to determine the Gittins indices for the possible branching projects which may arise. Then, if the value of the best available product is greater than the highest Gittins index of the available (sub-) projects, the agent stops experimenting and makes that product; else she works on the (sub-)project with the currently highest index, and continues.

The possible projects can be classified into four types: either a project contains just a terminal action (making a product), or it is a branching project of depth 1, 2, or 3 (corresponding to a development project, a research project, and a feasibility study respectively). These are illustrated in Figure 1.5. The rest of the analysis of this section concerns representative projects, and we adopt the convention that a representative project of type d corresponds to production if $d = 0$, and is a branching project of depth d if $d > 0$.[15] The *initial state* of a such project is the

[15]Subscripts on parameters, variables and functions, etc. will henceforth indicate the project depth and no longer the node, but when discussing generic properties we omit the subscript.

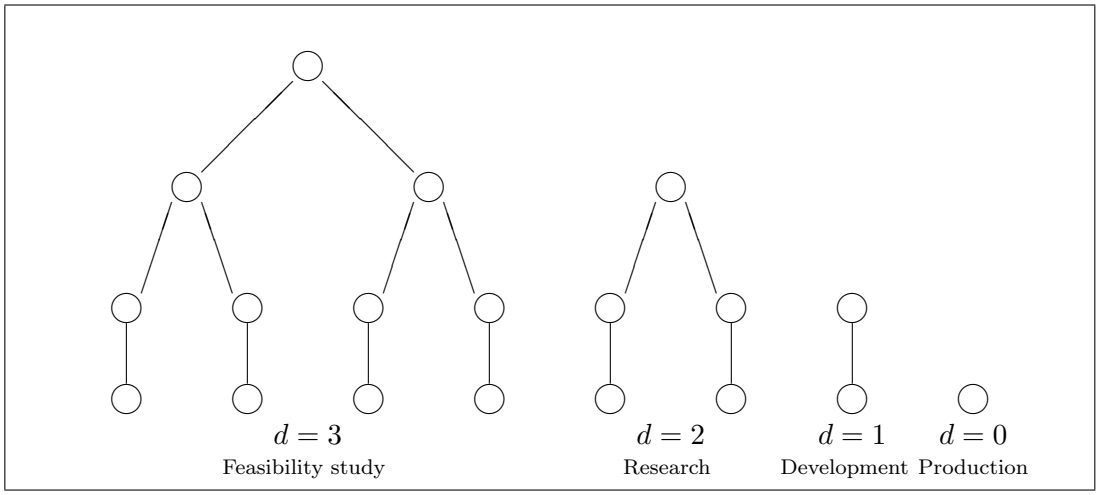


Figure 1.5: Four types of branching project

state when only the root action is available, and is a summary of everything known about the products which may emerge from that project. The sum of the signals received on taking actions which are ancestors of the root is such a summary, which we denote by y. Consider a project of type $d > 0$ and suppose that it is in its initial state y at time t. If the agent takes the root action then she learns z_d and updates the expected value of the products in the project accordingly. The root action can now be ignored, being no longer available, and the products can be considered as being in one of the γ (sub-)projects of type $d-1$, each of which is in its initial state $y + z_d$ at time $t + 1$.

To find the Gittins index for a project, consider the process which consists of just that project and a fall-back m, and let $\phi(m, y)$ denote the value of this process when the initial state of the project is y. Denote the Gittins index, or reservation price, of the project by $r(y)$. By definition, if $m > r(y)$ the agent stops with the fall-back m, otherwise she pays c to learn the increment z and then continues. Denoting the continuation value by $\tilde{\phi}(m, z + y)$, we have the general formula for the continuation region:

$$\phi(m, y) = -c + \mathrm{E}\Big[\tilde{\phi}(m, z + y) \;\; | \;\; y\Big].$$

As the Gittins index is the minimal fall-back which makes the agent indifferent between stopping and continuation, we see that $r(y) = \phi(r(y), y)$, so $r(y)$ satisfies:

$$r(y) = -c + \mathrm{E}\Big[\tilde{\phi}(r(y), z + y) \;\; | \;\; y\Big].$$

For the rest of the section we will make the following simplifying assumption.

Assumption

(a) there is no discounting, i.e. $\beta = 1$;

(b) the number of branches emanating from the root of any project of type $d > 0$ is the same, namely γ_{d-1}, with $\gamma_0 = 1$;

(c) the cost of visiting the root of any project of type $d > 0$ is the same, namely c_d;

(d) the signal z_d received at the root of any project of type $d > 0$ is independently drawn from the same continuous distribution with support $[a_d, b_d]$, CDF $G_d(\cdot)$ and pdf $g_d(\cdot)$.

It will transpire that $r(y) = r(0) + y$, which is intuitively plausible: if the agent is indifferent between an project with initial value y and a fall-back of $r(y)$, she will also be indifferent between that project with initial value 0 and a fall-back of $r(y) - y$.

The implication of the above remark, together with the assumption, is that the optimal policy in our example will be fully characterised by just four quantities, namely r_0, r_1, r_2 and r_3, the index for each of the four types of project when the initial state is zero. We now derive expressions for these.

Production

As we have assumed that production is costless and its value is known, in this case c is zero, z is the degenerate random variable equal to zero, and so the continuation pay-off is simply the larger of m and y, i.e. $\tilde{\phi}(m, z + y) = m \vee y$. So, subscripting the variables and functions by 0:

$$r_0(y) = r_0(y) \vee y$$

and the minimal $r_0(y)$ which satisfies this is clearly given by $r_0(y) = y$. For consistency with what follows, we define r_0 as $r_0(0)$, and then we have:

$$\begin{aligned} r_0 &= 0 \\ r_0(y) &= r_0 + y. \end{aligned}$$

Development

In the continuation region for *production* ($m \leq r_0 + y$):

$$\phi_0(m, y) = m \vee y$$

and indeed in general:

$$\phi_0(m, y) = m \vee y.$$

For development, we subscript the variables and functions by 1. If the agent reveals z_1 she will be facing a single production project, the value of which will be $\phi_0(m, z_1 + y)$. So, in the continuation region for *development*:

$$\begin{aligned}
\phi_1(m, y) &= -c_1 + \mathrm{E}\Big[m \vee (z_1 + y) \mid y\Big] \\
&= -c_1 + \int_{a_1}^{b_1} m \vee (z_1 + y)\, dG_1(z_1) \\
&= -c_1 + m + \int_{a_1}^{b_1} 0 \vee (z_1 + y - m)\, dG_1(z_1) \\
&= -c_1 + m + \int_{m-y}^{b_1} (z_1 + y - m)\, dG_1(z_1) \\
&= -c_1 + m + \int_{m-y}^{b_1} (1 - G_1(z_1))\, dz_1
\end{aligned}$$

the last line following from integrating by parts. So, from indifference:

$$\begin{aligned}
r_1(y) &= -c_1 + r_1(y) + \int_{r_1(y)-y}^{b_1} (1 - G_1(z_1))\, dz_1 \\
c_1 &= \int_{r_1(y)-y}^{b_1} (1 - G_1(z_1))\, dz_1.
\end{aligned}$$

This implicitly defines the value of $r_1(y) - y$ in terms of c_1 and the CDF $G_1(\cdot)$, and this value is therefore independent of y. As above, we define r_1 as $r_1(0)$, and then we have:

$$\begin{aligned}
c_1 &= \int_{r_1}^{b_1} (1 - G_1(z_1))\, dz_1 \\
r_1(y) &= r_1 + y.
\end{aligned}$$

Research

In the continuation region for *development* ($m \leq r_1 + y$):

$$\begin{aligned}
\phi_1(m, y) &= -c_1 + m + \int_{m-y}^{b_1} (1 - G_1(z_1))\, dz_1 \\
&= m + \int_{m-y}^{b_1} (1 - G_1(z_1))\, dz_1 - \int_{r_1}^{b_1} (1 - G_1(z_1))\, dz_1 \\
&= m + \int_{m-y}^{r_1} (1 - G_1(z_1))\, dz_1
\end{aligned}$$

and in general:

$$\phi_1(m,y) = m \ \vee \ \Big(m + \int_{m-y}^{r_1} (1 - G_1(z_1))\, dz_1\Big).$$

In the case of a research project, if the agent reveals z_2 she will be facing several development projects, the value of each of which will be $\phi_1(m, z_2 + y)$. Let $\Phi_1(M, y)$ denote the value of these γ_1 projects when the fall-back is M and the state of each of them is summarised by y. Using the formula given in Theorem 2.1, we have:

$$\Phi_1(M,y) = B - \int_M^B \Big(\frac{\partial}{\partial m}\phi_1(m,y)\Big)^{\gamma_1} dm$$

where B is the bound on the reward functions. In the stopping region ($m > r_1 + y$), the partial derivative is 1, otherwise, in the continuation region, $\partial\phi_1(m,y)/\partial m = G_1(m-y)$. Thus:

$$\Phi_1(M,y) = M \ \vee \ \Big(M + \int_{M-y}^{r_1} (1 - G_1(z_1)^{\gamma_1})\, dz_1\Big).$$

So, in the continuation region for the research project:

$$\begin{aligned}
\phi_2(m,y) &= -c_2 + \mathrm{E}\Big[\Phi_1(m, z_2 + y) \ \mid \ y\Big] \\
&= -c_2 + \int_{a_2}^{b_2} m \ \vee \ \Big(m + \int_{m-y-z_2}^{r_1} (1 - G_1(z_1)^{\gamma_1})\, dz_1\Big)\, dG_2(z_2) \\
&= -c_2 + m + \int_{a_2}^{b_2} 0 \ \vee \ \Big(\int_{m-y-z_2}^{r_1} (1 - G_1(z_1)^{\gamma_1})\, dz_1\Big)\, dG_2(z_2) \\
&= -c_2 + m + \int_{m-y-r_1}^{b_2} \Big(\int_{m-y-z_2}^{r_1} (1 - G_1(z_1)^{\gamma_1})\, dz_1\Big)\, dG_2(z_2) \\
&= -c_2 + m + \int_{m-y-r_1}^{b_2} \Big[1 - G_1(m-y-z_2)^{\gamma_1}\Big](1 - G_2(z_2))\, dz_2
\end{aligned}$$

the last line again following from integrating by parts. Again using $r_2(y) = \phi_2(r_2(y), y)$, we obtain:

$$c_2 = \int_{r_2(y)-y-r_1}^{b_2} \Big[1 - G_1(r_2(y) - y - z_2)^{\gamma_1}\Big](1 - G_2(z_2))\, dz_2$$

This time, it is not as obvious that this equation uniquely determines the value of $r_2(y) - y$. However, having observed that, say, an increase in $r_2(y) - y$ would decrease both the integrand and the range of integration whilst leaving the LHS unchanged, we conclude as before that $r_2(y) - y$ is independent of y and so we

define r_2 as $r_2(0)$ to give:

$$\begin{aligned} c_2 &= \int_{r_2-r_1}^{b_2} \Big[1 - G_1(r_2 - z_2)^{\gamma_1}\Big](1 - G_2(z_2))\, dz_2 \\ r_2(y) &= r_2 + y. \end{aligned}$$

Feasibility study

In the continuation region for *research* ($m \leq r_2 + y$):

$$\begin{aligned} \phi_2(m, y) &= -c_2 + m + \int_{m-y-r_1}^{b_2} \Big[1 - G_1(m - y - z_2)^{\gamma_1}\Big](1 - G_2(z_2))\, dz_2 \\ &= m + \int_{m-y-r_1}^{b_2} \Big[1 - G_1(m - y - z_2)^{\gamma_1}\Big](1 - G_2(z_2))\, dz_2 \\ &\quad - \int_{r_2-r_1}^{b_2} \Big[1 - G_1(r_2 - z_2)^{\gamma_1}\Big](1 - G_2(z_2))\, dz_2 \\ &= m + \int_{m-y-r_1}^{r_2-r_1} \Big[1 - G_1(r_2 - z_2)^{\gamma_1}\Big](1 - G_2(z_2))\, dz_2 \\ &\quad + \int_{m-y-r_1}^{b_2} \Big[G_1(r_2 - z_2)^{\gamma_1} - G_1(m - y - z_2)^{\gamma_1}\Big](1 - G_2(z_2))\, dz_2. \end{aligned}$$

The derivation of the Gittins index for a feasibility study follows the same steps as above for a research project. As the calculations are somewhat laborious (see Chapter 2, Appendix), we simply note that $r_3(y) = r_3 + y$, state the implicit formula for r_3, and collect the results together.

Reservation prices

0: $r_0 = 0$

1: $c_1 = \int_{r_1}^{b_1} (1 - G_1(z_1))\, dz_1$

2: $c_2 = \int_{r_2-r_1}^{b_2} \Big[1 - G_1(r_2 - z_2)^{\gamma_1}\Big](1 - G_2(z_2))\, dz_2$

3: $c_3 = \int_{r_3-r_2}^{b_3} \Big(1 - \Big[1 - \int_{r_3-z_3-r_1}^{b_2} (1 - G_1(r_3 - z_3 - z_2)^{\gamma_1}) g_2(z_2)\, dz_2\Big]^{\gamma_2}\Big)(1 - G_3(z_3))\, dz_3$

3.2 Implications of the Optimal Strategy

Much of the intuition underlying the determinants of the index and so of the following result is illustrated by considering how r_1, the index for a development project, depends on the 'riskiness' of the pay-offs. In a development project (with an initial value of zero) there are two actions: the root action is to observe a signal z_1, and its child is to make the product whose value is z_1. The Gittins index is given by the formula $c_1 = \int_{r_1}^{b_1} (1 - G_1(z_1))\, dz_1$. Notice that the Gittins index does not depend

on the distribution of low values of z_1, because when deciding how to proceed the agent always has the option, exercised if z_1 is low, of taking the fall-back rather than making the new product. The idea that the Gittins index depends just on the likelihood of high outcomes is captured by the result that r_1 increases if we consider a mean-preserving spread of the distribution of z_1.[16]

Result 3.1 *Let $H(\cdot)$ and $G(\cdot)$ be two CDFs such that $H(\cdot)$ is a mean-preserving spread of $G(\cdot)$ with the 'single-crossing property'. The Gittins index of the single stage project whose pay-off is distributed according to $H(\cdot)$ is greater than that of a similar project whose pay-off is distributed according to $G(\cdot)$.*[17]

PROOF: When H and G have the same mean:

$$\int_a^b (1 - H(z))\, dz = \int_a^b (1 - G(z))\, dz.$$

When $H(\cdot)$ is a spread of $G(\cdot)$ with the single-crossing property:

$$\int_a^x (H(z) - G(z))\, dz \geq 0$$

with equality at $x = a$ and $x = b$ and strict inequality for some $a < x < b$. Together,

$$\int_x^b (H(z) - G(z))\, dz \leq 0.$$

Denoting the two reservation prices by r_H and r_G, we have by definition:

$$c \;=\; \int_{r_H}^b (1 - H(z))\, dz = \int_{r_G}^b (1 - G(z))\, dz,$$

so

$$\begin{aligned} 0 \;&=\; \int_{r_H}^b (1 - H(z))\, dz - \int_{r_G}^b (1 - G(z))\, dz \\ &=\; \int_{r_H}^b (G(z) - H(z))\, dz - \int_{r_G}^{r_H} (1 - G(z))\, dz. \end{aligned}$$

The first integral is non-negative, and so $\int_{r_G}^{r_H} (1 - G(z))\, dz \geq 0$.

This implies that $r_H \geq r_G$, and if there is some difference in H and G towards the upper end of their support then the inequality is strict. ■

Thus if there is a choice between two development projects in which the expected value of the product from each project is the same, but with different variance, then it is optimal to do the more risky development first.

[16] This is another illustration of the difference between the Gittins index of a project and its expected value.

[17] *This point is explored in Weitzman [1979].*

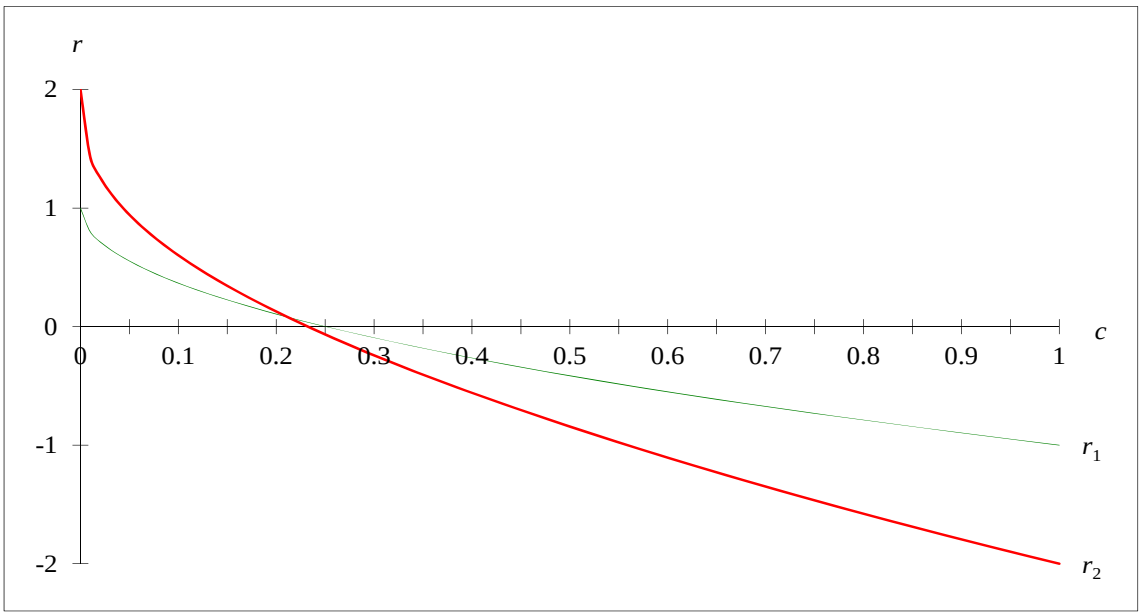


Figure 1.6: Reservation prices *v.* cost

Example 3.1 *Reservation prices as a function of cost*

The above result can be used to understand the relative behaviour of the Gittins indexes r_1 and r_2 as the cost of experimentation increases. For the case with two-way branching, equal costs of research & development, and where the distribution is uniform on $[-1, 1]$, the reservation prices vary with costs as shown in Figure 1.6.

The indexes r_1 and r_2 are calculated assuming that the initial states of the projects are zero.[18] Also note that the expected value of any signal is zero. Since the value of a product in a project is the sum of that project's initial state and the signals about the product that are subsequently observed, then initially the expected value of any product in both the research project and the pure development project is zero. However, the values of products in the research project have a higher variance. When the cost of search is low, this difference in the variance is the main consideration, and as we would expect from Result 3.1, the Gittins index for research is higher than that for development. As the

[18] It is easy to show that r_1 satisfies $c_1 = [(1 - r_1)/2]^2$, giving $r_1 = 1 - 2\sqrt{c}_1$ for $0 \leq c_1 \leq 1$.
Determining r_2 is a little more complicated. For $0 \leq c_1, c_2 \leq 1$, it is the positive root which is less than 2 of

$$m^4 - 24m^2 + 32(2 - 3c_1 + c_1\sqrt{c}_1)m - 48(1 - 4c_1 + 4c_1\sqrt{c}_1 - c_1^2) + 96c_2 = 0$$

and it is the negative root which is greater than -2 of

$$-24m^2 + 32(2 - 3c_1 + c_1\sqrt{c}_1)m - 48(1 - 4c_1 + 4c_1\sqrt{c}_1 - c_1^2) + 96c_2 = 0.$$

search cost rises, however, a new consideration becomes increasingly important: the agent must spend more before production if the product is at the end of a research project than if it is in a development project. Thus as the cost rises, the Gittins index for development becomes higher than that for research. ♢

The main focus of this section is on how branching affects the way that agents pursue R&D. The example above shows that as costs rise, the balance tips in favour of pursuing development before engaging in more research, and this remains qualitatively the case if we allow the amount of branching to vary. In the example below, we shall see that as branching increases the agent tends to do more initial research before embarking on any development. First, note that the expected value of a project consisting of γ_1 identical development opportunities is $r_1 - \int_{a_1}^{r_1} G_1(z_1)^{\gamma_1}\, dz_1$, which is increasing in γ_1, the number of branches from their common research parent.[19] Next, consider the effect of the amount of branching on the Gittins index for research (it has no effect on the Gittins index for development).

Result 3.2 *As the amount of branching increases, r_2 increases and r_1 is unchanged.*

PROOF: The expression giving r_2 implicitly is

$$c_2 = \int_{r_2 - r_1}^{b_2} \left[1 - G_1(r_2 - z_2)^{\gamma_1}\right](1 - G_2(z_2))\, dz_2$$

If we hold r_2 fixed and increase γ_1, then the right-hand side increases. To restore the equality with c_2, we must increase r_2 thereby decreasing the range of integration and also the term $[1 - G_1(r_2 - z_2)^{\gamma_1}]$.

The expression for r_1 is independent of the amount of branching. ■

Now, what is the probability that, having explored one research avenue, the agent prefers to explore a second research avenue before pursuing any development of the first? Assume, without loss of generality, that the signal received from the feasibility study was zero. If the signal received from the first piece of research is z, then the Gittins index for developments of that research is $r_1 + z$. Thus the agent will undertake a second piece of research if $r_2 > z + r_1$ so that the probability of doing the second piece of research first is given by $\Pr(r_2 > z + r_1)$, which is just $G(r_2 - r_1)$. As we would expect, this is increasing in the reservation price of research

[19] This leads to the final illustration of the difference between the reservation price for a project and its expected value. The reservation price for a project consisting of γ_1 identical development opportunities is simply the reservation price for just *one* development opportunity, namely r_1. This is strictly greater than the project's expected value noted above, which approaches the reservation price as the amount of branching tends to ∞.

and decreasing in the reservation price of development. The final example presented here is a direct consequence of Result 3.2.

Example 3.2 *As the number of ways of developing a single piece of research increases, the agent is more likely to do a second piece of research before pursuing any development of the first.*

The intuition behind this is that the larger the number of development opportunities from a single research avenue, the higher are the expected rewards from after the development phase, and so it becomes more attractive to learn about these expected rewards before pursuing existing development opportunities. ♢

4 Conclusion

The central innovation of the chapter is the introduction of a sequential search process which can be represented as a family of trees, and the central theoretical result is that the optimal action to take in this process is given by a Gittins index policy. This result extends the existing work on multi-armed bandits in the economics literature in two important ways. In existing models, either projects are fully independent and the Gittins index policy is optimal, or they are not independent and the models have no such simplifying result. In our process the stochastic specification means that actions can have correlated rewards, so that independence is relaxed, yet the index policy remains optimal.

The second generalisation is that in existing multi-armed bandit models there is just one action available in each project in any one period, whereas in our process, the agent constantly faces choices about the direction in which to advance a project. The technical device which allows us to do this, while maintaining the result that the Gittins index policy identifies the optimal *action* and not just the optimal *project*, is to recognise that the way that actions are grouped into projects need not be the same in every period.

The final part of the chapter turns to economic applications. The representation of the process as a family of trees reflects the notion of precedence: some actions follow on from others; and it gives a measure of the diversity of rewards: close rewards have a nearer common ancestor than distant ones. The process also generates the feature that close rewards are more highly correlated than distant ones. This structure is clearly a natural one within which to study R&D and technological change[20]

[20]Vega-Redondo [1993] independently develops a similar model, though there the author's focus

and we investigate a very simple model of R&D in order to illustrate the main technical result. We find that as costs rise the agent expects to pursue development before engaging in more research, but that as the amount of branching increases, the agent expects to do more research before embarking on any development.

There are several ways in which this work could be extended. As mentioned in the introduction, modelling R&D as searching a branching structure provides a means of investigating the diversity of products that are developed and marketed, and how this depends on the nature of competition in R&D. Branching projects also provide a framework within which to examine the dual role of patents as not simply conferring monopoly rights over some products, but simultaneously revealing information about related products not covered by the patent. Some of these topics will be addressed in follow-up research.

is on industry turnover rather than optimal search. Furusawa [1994] also employs a branching structure to aid a game-theoretic analysis of the costs and benefits of Research Joint Ventures.

Appendix

A Optimality of the Gittins index policy for simple bandit processes

We give an outline of the proof of the optimality of the Gittins index policy for multi-armed bandits; it is essentially from Whittle [1982], and also used in Berry and Fristedt [1985]. It is included here for accessibility, and contains some notational changes and expository material due to the present authors.

There are N projects[21] and in each discrete period you can work on only one project. The state of project k at time t is denoted by $x_k(t)$, and the project engaged at time t is denoted by $k(t)$. The state variable at time t is written as $x(t) = \langle x_1(t), x_2(t), \ldots, x_N(t)\rangle$, and the information at time t, namely *past and current* states and *past* actions, is written as $I(t)$. If project k is engaged at time t then you get an immediate expected reward of $R_k(x_k(t))$. Rewards are additive and discounted in time by a factor β. States of unengaged projects do not change and the state of the engaged project changes by a Markov transition rule: if $k(t) \neq k$ then $x_k(t+1) = x_k(t)$, and if $k(t) = k$ then the value of $x_k(t+1)$ is conditioned only by k & $x_k(t)$.

Assume that rewards are uniformly bounded:

$$-\infty < -B(1-\beta) \leq R_k(x) \leq B(1-\beta) < \infty.$$

Writing $R(t)$ for the reward $R_{k(t)}(x_{k(t)}(t))$ realised at time t, the total discounted reward is then $\sum_0^\infty \beta^t R(t)$ with a maximal expected reward $F(x)$ over feasible policies π given by:

$$F(x(0)) = \sup_\pi \mathrm{E}_\pi\Big[\sum_0^\infty \beta^t R(t) \;\mid\; I(0)\Big].$$

F will be the unique bounded solution to the dynamic programming equation:

$$F = \max_k L_k F$$

where L_k is the one-step operator if k is the project engaged:

$$L_k F(x) = R_k(x_k) + \beta\,\mathrm{E}\Big[F(x(t+1)) \;\mid\; x(t) = x, k(t) = k\Big].$$

Introduce a fall-back M, where the option of taking the fall-back remains open at all

[21] Here, we are dealing with single-action projects. At any stage, each of the fixed number of projects has a single action (i.e. there is no branching) so that the notions of engaging a project and selecting an action are interchangeable.

times. The maximal expected reward of the modified process, conditional on $x(0) = x$, is $\Phi(M, x)$ and solves

$$\Phi = M \vee \max_k L_k \Phi. \tag{A.1}$$

Let $\phi_k(m, x_k)$ be the analogue of $\Phi(M, x)$ when only project k is available; ϕ_k solves

$$\phi_k = m \vee L_k \phi_k. \tag{A.2}$$

(L_k changes only x_k, so $L_k \phi_k$ is well-defined.)

The *Gittins index*, denoted by $m_k(x_k)$, is the infimal value of m such that $m = \phi_k(m, x_k)$, namely the alternatives of stopping with m_k and of continuing project k (with the option of taking the fall-back staying open) are equitable, and so $m_k = L_k \phi_k$.

It is fairly easy to show that $\Phi(M, x)$, as a function of M, is non-decreasing & convex (convexity following from the fact that we are dealing with the supremum of expressions which are linear in M), and that $\Phi(M, x) = M$ when $M \geq B$. Also $\Phi(M, x) = F(x)$ when $M \leq -B$.

Similarly, $\phi_k(m, x_k)$, as a function of m, is non-decreasing & convex, and $\phi_k(m, x_k) = m$ when m is large, certainly if $m \geq B$, and more precisely for $m \geq m_k$, so $m_k \leq B$. Note that, since $\phi_k(m, x_k)$, as a function of m, is convex, the derivative $\partial \phi_j(m, x_j)/\partial m$ exists almost everywhere.

We "guess" the form of the value function:

$$\Theta(M, x) = B - \int_M^B \prod_k \frac{\partial \phi_k(m, x_k)}{\partial m} \, dm,$$

and proceed to verify it by showing two things:

- Θ satisfies (A.1), that is $\Theta = M \vee \max_k L_k \Theta$;
- the action recommended by the Gittins index maximises the RHS of the above equation, i.e. when $M > \max_k L_k \Theta$ it selects the fall-back, and when $M < \max_k L_k \Theta$ it selects the project which maximises $L_k \Theta$.

So, define $P_k(m, x)$ and $m_{\neg k}$ by

$$\begin{aligned} P_k(m, x) &= \prod_{j \neq k} \frac{\partial \phi_j(m, x_j)}{\partial m} \\ \text{and } m_{\neg k} &= \max_{j \neq k} m_j. \end{aligned}$$

$P_k(m, x)$, as a function of m, is non-negative and non-decreasing, and $P_k(m, x) = 1$ for $m \geq m_{\neg k}$. (These follow directly from the properties of ϕ_j.) Note that

$$d_m P_k(m, x) \geq 0 \ \forall \ m, \text{ and } d_m P_k(m, x) = 0 \text{ for } m \geq m_{\neg k}.$$

Rewrite $\Theta(M,x)$ as

$$\Theta(M,x) = B - \int_M^B \frac{\partial \phi_k(m,x_k)}{\partial m} P_k(m,x)\, dm$$

and use integration by parts to obtain

$$\begin{aligned}\Theta(M,x) &= B - \Big[\phi_k(m,x_k)P_k(m,x)\Big]_M^B + \int_M^B \phi_k(m,x_k)\, d_m P_k(m,x) \\ &= \phi_k(M,x_k)P_k(M,x) + \int_M^B \phi_k(m,x_k)\, d_m P_k(m,x)\end{aligned}$$

noting that $\phi_k(B,x_k) = B$ because $m_k \leq B$, and $P_k(B,x) = 1$ because $m_{\neg k} \leq B$. Also, $d_m P_k(m,x) = 0$ when $m \geq B$, so we can amend the range of integration:

$$\Theta(M,x) = \phi_k(M,x_k)P_k(M,x) + \int_M^\infty \phi_k(m,x_k)\, d_m P_k(m,x). \tag{A.3}$$

Now fix x, so we can focus on the dependence of various function on m or M. We want to show that:

$$\begin{aligned}\Theta(M) &\geq M \text{ for any } M, && \text{(A.4)}\\ \text{and } \Theta(M) &= M \text{ iff } M \geq \max_j m_j; && \text{(A.5)}\\ \text{and that } \Theta(M) &\geq L_k\Theta(M) \text{ for any } M, && \text{(A.6)}\\ \text{and } \Theta(M) &= L_k\Theta(M) \text{ iff } m_k = \max_j m_j \text{ and } M \leq m_k. && \text{(A.7)}\end{aligned}$$

• (A.5) 'if' & part of (A.4):

Consider $M \geq \max_j m_j$. In this case, $\phi_k(M) = M$, $P_k(M) = 1$; and $d_m P_k(m) = 0$ for $m \geq M$. So from (A.3):

$$\Theta(M) = M.$$

• (A.5) 'only if' & rest of (A.4):

Consider $M < \max_j m_j$. Let $k = \arg\max_j m_j$. So $M < m_k$, and we have $\phi_k(M) > M$. When $M \leq m < m_k$, $\phi_k(m) > M$, and when $m \geq m_k$, $d_m P_k(m) = 0$. So from (A.3):

$$\begin{aligned}\Theta(M) \;>\; M\left(P_k(M) + \int_M^{m_k} d_m P_k(m)\right) &= M P_k(m_k) \\ &= M, \text{ because } P_k(m_k) = 1.\end{aligned}$$

So from (A.4) and (A.5):

$$\begin{aligned}M < \max_j m_j &\Rightarrow \Theta(M) > M \\ M = \max_j m_j &\Rightarrow \Theta(M) = M \\ M > \max_j m_j &\Rightarrow \Theta(M) = M.\end{aligned} \tag{A.8}$$

Now, define $\delta_k(m, x_k)$ by

$$\delta_k(m, x_k) = \phi_k(m, x_k) - L_k\phi_k(m, x_k).$$

Fixing x again, note that $\delta_k(m) \geq 0 \; \forall \; m,$ and $\delta_k(m) = 0$ for $m \leq m_k$ and that

$$\Theta(M) - L_k\Theta(M) = \delta_k(M)P_k(M) + \int_M^\infty \delta_k(m)\, d_m P_k(m) \tag{A.9}$$

which follows from applying the one-step operator L_k to each side of (A.3), subtracting the result from (A.3), and applying the definition of δ_k to the RHS.

• (A.6):

$\delta_k(m) \geq 0$, and $P_k(m)$ is non-negative and non-decreasing, so from (A.9) we have

$$\Theta(M) \geq L_k\Theta(M).$$

• (A.7):

When $m_k \geq M$, $\delta_k(M) = 0$, so the first term on the RHS of (A.9) is 0.
When $m_k \geq M$ and $m_k = \max_j m_j$, the integral on the RHS of (A.9) is 0, because either $\delta_k(m) = 0$, or $d_m P_k(m) = 0$, or both. So we have

$$\Theta(M) = L_k\Theta(M).$$

(If either $m_k < M$ or $m_k < m_{\neg k}$, then at least one term on the RHS of (A.9) is positive.)

Now using (A.6) & (A.7) with the implications from (A.8), and with $k = \arg\max_j m_j$:

$$\begin{aligned} M < \max_j m_j = m_k \;&\Rightarrow\; \Theta(M) = L_k\Theta(M) \text{ and } \Theta(M) > L_{\neg k}\Theta(M) \\ &\Rightarrow\; \max_j L_j\Theta = L_k\Theta = \Theta > M, \text{ so } \{M \vee \max_j L_j\Theta\} = L_k\Theta; \end{aligned}$$

$$\begin{aligned} M = \max_j m_j = m_k \;&\Rightarrow\; \Theta(M) = L_k\Theta(M) \text{ and } \Theta(M) > L_{\neg k}\Theta(M) \\ &\Rightarrow\; \max_j L_j\Theta = L_k\Theta = \Theta = M, \text{ so } \{M \vee \max_j L_j\Theta\} = M \equiv L_k\Theta; \end{aligned}$$

$$\begin{aligned} M > \max_j m_j = m_k \;&\Rightarrow\; \Theta(M) > L_k\Theta(M) \text{ and } \Theta(M) > L_{\neg k}\Theta(M) \\ &\Rightarrow\; \max_j L_j\Theta < \Theta = M, \text{ so } \{M \vee \max_j L_j\Theta\} = M. \end{aligned}$$

So Θ satisfies (A.1), that is $\Theta = M \vee \max_j L_j\Theta$, and the Gittins index policy is optimal. ■

Thus, $\Theta = \Phi$ and the following identity holds:

$$\Phi(M, x) = B - \int_M^B \prod_k \frac{\partial \phi_k(m, x_k)}{\partial m}\, dm. \tag{A.10}$$

Whittle [1982, §9] indicates that the proof can be modified to incorporate variable length project stages.

Assume that when one engages project k in state x_k then one is committed to it for a *stage* of length $s = s(k, x_k)$. We shall suppose that s and $x_k(t+s)$ are conditioned only by k and x_k, and not by t. The dynamic programming equations become recursions between discrete *stages* instead of between discrete *periods*, and we modify the definition of the one-step operator L_k:

$$L_k F(x) = R_k(x_k) + \mathrm{E}\Big[\beta^s F(x(t+s)) \;\mid\; x(t) = x, k(t) = k\Big]$$

where $R_k(x_k)$ is now the reward from the *stage* starting from state x_k.

The single project return $\phi_k(m, x_k)$ defined in (A.2) is now in terms of the modified L_k, and the identity (A.10) after the end of the proof of the main result still holds between Φ and the ϕ_k; the Gittins index policy is optimal. ∎

B Optimality of the Gittins index policy for bandit super-processes

We now show how Whittle's proof (outlined above) of the optimality of the Gittins index policy for simple processes (consisting of single-action projects) can be generalised to cover super-processes (consisting of multi-action projects).

Remember, a super-process is one in which, after a project has been chosen, there is a further decision to be made as to how to proceed, and this affects both the reward and the state transition of the chosen project. The proof of the optimality of the Gittins index policy for super-processes fails except in one special case, which is when the following condition holds: *the optimal subsidiary decision as to how to proceed with the chosen project is independent of the size of the fall-back.* (In other words, if a project is the only one available then your optimal action does not change when the fall-back varies over the range in which you prefer to continue with the project.) The proof below that this condition is sufficient elaborates on that in Whittle [1980]. That this condition is also necessary can be found in Glazebrook [1982].

There are N projects, each project having possibly more than one available action when in a given state, and in each discrete period you can take only one action and thus work on only one project. The state of project k at time t is denoted by $x_k(t)$ and the state variable at time t is written as $x(t) = \langle x_1(t), x_2(t), \ldots, x_N(t)\rangle$. The set of available actions for project k in state x_k is denoted by $U_k(x_k)$, and the set of all available actions is the union over k of these, denoted by $U(x)$. Let $\kappa(\cdot)$ be the indicator function mapping available actions to projects, i.e. $\kappa(u) = k$ for $u \in U_k$. The action taken at time t is denoted by $u(t)$, and thus the project engaged at time t is $\kappa(u(t))$. If action u is taken at time t then you get an immediate expected reward of $R_{\kappa(u)}(x_{\kappa(u)}(t), u)$. Rewards are additive and discounted in time by a factor β. States of unengaged projects do not change and the state of the engaged project changes by a Markov transition rule: if $\kappa(u(t)) \neq k$ then $x_k(t+1) = x_k(t)$, and if $\kappa(u(t)) = k$ then the value of $x_k(t+1)$ is conditioned only by $u(t)$, k & $x_k(t)$.

Continue to assume that rewards are uniformly bounded:

$$-\infty < -B(1-\beta) \leq R_k(x,u) \leq B(1-\beta) < \infty.$$

When m is the available fall-back, $\phi_k(m, x_k)$ now solves

$$\phi_k = m \ \vee \ \sup_{u \in U_k} L_{k,u}\phi_k \tag{B.1}$$

where

$$L_{\kappa(u),u}\Phi(M,x) = R_{\kappa(u)}(x_{\kappa(u)}, u) + \beta\, \mathrm{E}\Big[\Phi(M, x(t+1)) \ \mid \ M, x(t) = x, u(t) = u\Big].$$

As usual, the *Gittins index* of project k, denoted by $m_k(x_k)$, is the infimal value of m such that $m = \phi_k(m, x_k)$, namely the alternatives of stopping with m_k and of embarking on project k (with the option of taking the fall-back staying open) are equitable, and so $m_k = \sup_{u \in U_k} L_{k,u}\phi_k$.

$\Theta(M,x)$ is defined as before, and we still have (A.3):

$$\Theta(M,x) = \phi_k(M,x_k)P_k(M,x) + \int_M^\infty \phi_k(m,x_k)\, d_m P_k(m,x)$$

so, having fixed x, the following ((A.4) & (A.5)) still hold:

$$\begin{aligned} \Theta(M) &\geq M \text{ for any } M, \\ \text{and } \Theta(M) &= M \text{ iff } M \geq \max_j m_j. \end{aligned}$$

The function $\delta(\cdot)$ is now action-specific not merely project-specific, so, for $u \in U_k$, define

$$\delta_{k,u}(m, x_k) = \phi_k(m, x_k) - L_{k,u}\phi_k(m, x_k).$$

Fixing x as before, to focus on m or M, note that

$$\Theta(M) - L_{k,u}\Theta(M) = \delta_{k,u}(M)P_k(M) + \int_M^\infty \delta_{k,u}(m)\, d_m P_k(m)$$

$$\begin{aligned} \text{so } \Theta(M) \;-\; & \sup_{u\in U_k} L_{k,u}\Theta(M) \\ & = \inf_{u\in U_k} \Big(\delta_{k,u}(M)P_k(M) + \int_M^\infty \delta_{k,u}(m)\, d_m P_k(m)\Big). \end{aligned} \tag{B.2}$$

We want to show that:

$$\Theta(M) \;\geq\; \sup_{u\in U_k} L_k\Theta(M) \text{ for any } M, \tag{B.3}$$

$$\text{and } \Theta(M) \;=\; \sup_{u\in U_k} L_k\Theta(M) \text{ iff } m_k = \max_j m_j \text{ and } M \leq m_k. \tag{B.4}$$

It is still the case that, for *any* $u \in U_k$, $\delta_{k,u}(m) \geq 0$ for all m, so inequality (B.3) still holds, and *if* we are able to assert that, for *some* $u \in U_k$, $\delta_{k,u}(m) = 0$ for $m \leq m_k$, then equality (B.4) also holds, by considering the RHS of (B.2). The assertion that such an action $u \in U_k$ exists is the same as saying that in the continuation region for the single project the optimal action is unique. ■

However, if there is not a unique optimal action $u \in U_k$ when $m \leq m_k$, then the RHS of (B.2) might be strictly positive for some $M \leq m_k$, in which case equality (B.4) would not hold, and the remainder of the proof would not go through.[22] To see this, suppose that a switch of actions occurs when the fall-back is $\hat{m}$, i.e. when m is such that $m \leq \hat{m}$ it is optimal to take action u', and when m is such that $\hat{m} \leq m \leq m_k$ it is optimal to take action u''. For action u' this implies that $\delta_{k,u'}(m) = 0$ when $m \leq \hat{m}$, & $\delta_{k,u'}(m) > 0$ when $\hat{m} < m \leq m_k$, and for action u'' this implies that $\delta_{k,u''}(m) > 0$ when $m < \hat{m}$, & $\delta_{k,u''}(m) = 0$ when $\hat{m} \leq m \leq m_k$. Consider $M < \hat{m}$, and suppose that the other projects under consideration are such that $P_k(M) > 0$ and $d_m P_k(m) > 0$ for $M \leq m < m_k$. Looking at the RHS of (B.2) for the two actions in turn we see that (a) the first term is zero because $\delta_{k,u'}(M) = 0$, but the integral is non-zero because neither $\delta_{k,u'}(m)$ nor $d_m P_k(m)$ is zero over $[\hat{m}, m_k]$, and (b) $\delta_{k,u''}(M) > 0$ and also the integral is non-zero (over $[M, \hat{m}]$). So the expression in parentheses on the RHS of (B.2) is strictly positive for either action, hence the infimum over the two actions is positive.

As in Appendix A, when the number of periods required to complete an action in a

[22] As an informal example of the second condition failing, consider a project with two actions: one leads to a state with a low mean value and a high variance; the other one leads to a state with a high mean value and a low variance. *Taking either action renders the other unavailable.* When the fall-back is high enough, it is optimal to take it. When the fall-back is lowered, it becomes optimal to take the more risky action, because if a poor outcome is realised there is always the fall-back. However, as the fall-back is lowered even further, it is no longer a good enough guarantee and so the optimal action switches to the less risky one.

project is different for different actions and different projects, the definition of the action of L_k can be suitably modified so that the above result remains valid.

Assume that when one takes action u in project $k = \kappa(u)$ in state x_k then one is committed to it for a *stage* of length $s = s(k, x_k, u)$. We shall suppose that s and $x_k(t+s)$ are conditioned only by k, x_k, and u, and not by t. The definition of the one-step operator L_k becomes:

$$L_{\kappa(u),u}F(x) = R_{\kappa(u)}(x_{\kappa(u)}, u) + \mathrm{E}[\beta^s F(x(t+s)) \;\mid\; x(t) = x, u(t) = u]$$

where $R_{\kappa(u)}(x_{\kappa(u)}, u)$ is now the reward from the *stage* starting from state x_k when action u is taken.

As before, the single project return $\phi_k(m, x_k)$ is now defined in terms of the modified L_k, and the identity (A.10) still holds between Φ and the ϕ_k; the Gittins index policy is optimal. ■

Chapter 2

Young Turks and Old Stagers

Introduction

There is a "stylised fact" that young inexperienced workers are attracted by jobs and occupations that exhibit risky returns, where the average return is on the low side and where turnover is high. On the other hand, older more experienced workers will have settled into jobs in occupations where turnover is lower, returns are on average higher and are more bunched around that mean.[1] One can put forward a variety of explanations for this behaviour: the young are impetuous, they have a presumption of success, they are overoptimistic and have unrealistic expectations, and so on. The aim of this chapter is to offer an alternative explanation, in which such behaviour is the rational outcome of an optimising agent.

In the model presented here, a job is treated as an *experience* good, that is, the agent finds out about the value to him of a particular job only after being hired. However, before looking for a job in a particular occupation (which is costly), the agent must first become qualified for that sector/ profession/ trade, which also has an associated cost and which is also an *experience* good – the training stage reveals the agent's general aptitude for jobs in the given occupation. So, there are precedence constraints, and costly information revelation at each stage; also, the returns to the agent are correlated within any occupation.

Despite this correlation, it can be shown that the agent is facing a multi-armed bandit problem and a simple 'reservation value' rule applies. By comparing various reservation values, we can determine the optimal behaviour of the agent, i.e. which

[1]Some of this is captured in the following extract from a study by Warwick Business School, reported in *The Independent*, Wednesday 15-Jan-97: "Wise, experienced 50 to 55-year-olds are more likely to survive in business than young, thrusting would-be entrepreneurs in their early twenties."

occupation to acquire skills in first, when to change jobs, whether or not to retrain for a different occupation, and so on. Further, we can address questions such as: How many jobs will the typical agent try before settling down? What return stream can the agent expect to end up with? What is the probability of the agent finding the most suitable job in a given occupation? It will transpire that new entrants to the job market (the 'Young Turks') will rationally launch themselves into professions in which the returns are more risky, whereas those with more experience (the 'Old Stagers') will have found themselves jobs that are suitable enough so that it is not worth their while to look for an alternative.

The chapter is organised as follows. In the next section, we introduce the general model. Then, in Section 2, we describe the optimal policy of the agent, and the implications for an agent who acts according to the reservation value rule. Section 3 is devoted to job turnover, expected returns, and related issues. The final section concludes, and indicates possible extensions. Technical derivations can be found in the appendix.

Related literature

The majority of models of Job Matching have treated jobs as *inspection* goods, where the value of the match is revealed prior to the match being proposed. (See, for example, Diamond [1982] and Pissarides [1990].) A major concern there is equilibrium wage determination, unemployment and unemployment duration, and how these vary with the business cycle. Here, we consider an unchanging search environment, and so will have little or nothing to say on those issues.

Other matching models of job search which treat jobs as experience goods are to be found in Jovanovic [1979], Miller [1984], and Felli & Harris [1996]. Jovanovic is concerned with wage determination in equilibrium and with exploring the relationship between tenure and turnover. Felli & Harris are also concerned with wage dynamics and turnover, and address the role of firm-specific human capital. Here, we treat wages as given, just 'out there' waiting to be discovered by the agent, and there is no strategic interaction unlike in Felli & Harris where two firms repeatedly compete for the services of a single worker. This chapter is more closely related to Miller's work in allowing for different job types, but with a less *ad hoc* notion of occupation. Miller's definition of 'occupation' is a collection of job prospects which are *ex ante* identical; all that we require is that the returns of jobs in any given occupation have a common component. If we were to adopt Miller's definition, and also his assumption of costless search and no prerequisite training, the analysis would be somewhat simpler and the results stronger (see Section 2.2).

With regard to the Learning literature, if the agent were to be constrained by having to choose a job from a specific occupation (for which he had already become qualified) and if the value of the match were perfectly revealed in the first period, then the agent would be facing Pandora's problem analysed in Weitzman [1979]. The more general underlying model of a sequential search process with precedence constraints was introduced in the previous chapter. It was shown there that, under certain conditions which obtain here, an index policy (specifically the Gittins index policy) recommends the action which is optimal, *despite* the returns to various actions being correlated. In this chapter, we extend that analysis to incorporate discounting and 'earn-as-you-go', and particularise the model to one of job search and occupational choice.

1 The Model

The general model of a sequential search process with precedence constraints can best be pictured as a collection of trees (each with the root at the top) in which you can explore a node only when you have already explored its parent (except for a root itself, of course, which by definition has no parent). Associated with any node is the cost of (or the reward from) exploring it, together with information about other nodes, for instance the returns available there.

To fix ideas, consider the choices available (now and in the future) as depicted in Figure 2.1. There are three occupations, each represented by a tree. The one on the left has 3 job opportunities in it, and the other two each have 2 jobs.

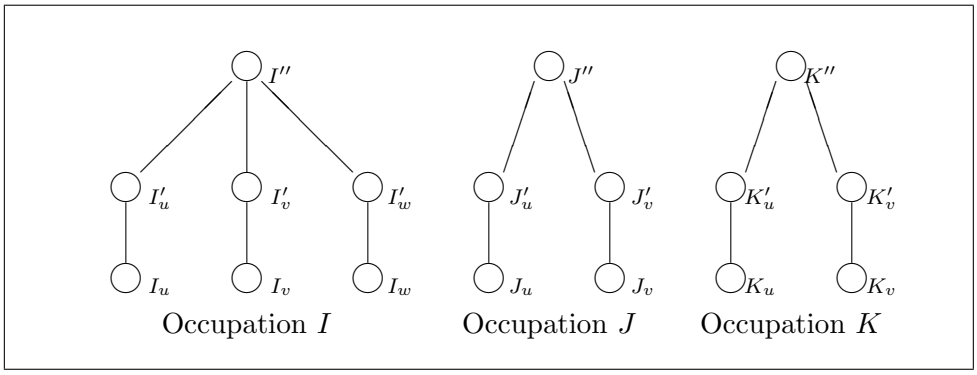


Figure 2.1: Three occupations

Precedence constraints and Information

At the outset, the agent is not qualified to work in any of the occupations, but say, for example, that he chooses occupation I. At some cost, the agent acquires the skills necessary to get any job in this occupation and he also learns about the expected returns from those jobs, that is, he explores node I''. Now, he can either find a job in this occupation (i.e. explore one of the nodes $I'_u \dots I'_w$), or retrain for a different occupation (i.e. explore either J'' or K''), the choice depending on what he learnt about his expected returns. Let us say that he takes the former action and explores I'_u. Then in this case, he will pay another one-off cost and his return in this job will be revealed[2] and received, and now he has three choices: (1) stay in this job (i.e. explore I_u) and get the known return forever;[3] (2) try to find a different job in this occupation (i.e. explore either I'_v or I'_w) with as yet unknown return; (3) train for a different occupation (i.e. explore either J'' or K'') in which the expected returns are not even known. And so on.

Note that training for one occupation (i.e. exploring node I'', for example), tells the agent nothing about his general aptitude for jobs in the other occupations. Further, subsequently getting a particular job in this occupation (i.e. exploring node I'_u, for example), tells the agent nothing about what his particular return would be in the other jobs in this occupation, and certainly not about jobs in other occupations.

Costs, Returns and Correlation

Consider the situation when the agent has learnt the return from working in a particular job in a given occupation, and that the return stream has a net present value (NPV) of $w = y + z$ say. Then the agent can stay in this job forever and receive $(1 - \beta)w$ each period at no cost, where β is the discount factor.

The return stream w has two components, so consider the situation when the agent is looking for a job in a given occupation. As a result of becoming qualified, he has learnt the common component, with NPV y say, that the returns to him from jobs in this occupation have. An arbitrary job will cost say c_1 to find, and it is only then that the NPV of his actual return stream in this particular job is revealed, that is he gets $-c_1 + (1 - \beta)(y + z)$ in the first period of employment, where z is drawn

[2]For simplicity, we assume that the return is perfectly revealed in the first period of work. We shall indicate how things would change if this were not the case.

[3]Think of node I_u as representing the net present value of the return stream from this job. As such, it is shorthand for an infinitely long non-branching chain of nodes, each of which represents the per-period return.

from a probability distribution with CDF $G_1(\cdot)$, pdf $g_1(\cdot)$, and support $[a_1, b_1]$,[4] and he receives $(1-\beta)(y+z)$ in each subsequent period that he works in this job.

Finally consider the situation before the agent has become qualified. The training for an arbitrary occupation will cost say c_2, and it is at this stage that he finds out about how suited he is to jobs in this occupation, that is, he learns the common component with NPV y which is drawn from a probability distribution, CDF $G_2(\cdot)$, pdf $g_2(\cdot)$, and support $[a_2, b_2]$.

We assume that once the agent has become qualified for an occupation, jobs within that occupation are available to him in any future period, even if he has retrained for a different occupation and/or tried a job in a different occupation in the meantime. Also, once the agent has found a job, then, if he quits, he may go back to that job in any future period without paying the search cost again.

Note that the returns from working in jobs within any given occupation are correlated – they share a common component which is revealed at the earlier qualification/ training stage; but the returns from working in jobs in different occupations are not correlated.[5]

The agent's problem

More formally, we represent an occupation by an out-tree which can have infinite depth but only a finite number of branches emanating from any node.[6] A node has either been explored or not; apart from a root node, any unexplored node can be explored only once its parent has been explored; once a node has been explored, there is no need, and indeed no possibility, to explore that node again. Associated with the exploration of a node is a reward whose expectation is bounded (and which may be positive or negative, i.e. a cost), and a signal containing information only about descendant nodes.[7] The rewards and signals do not depend on the time at which the action is taken. Time is discrete, and rewards are additive and discounted

[4]The subscript 1 indicates that parameters, variables and functions, etc., refer to the job search level. A subscript 2 will refer to the occupation choice level.

[5]We could of course precede the whole process described above by a requirement that the agent must first obtain some basic education, at which point his aptitude for work in general would be revealed. This is like preceding nodes I'', J'' and K'' by a new common root node. In this case, the agent's returns *across* occupations would also be correlated.

[6]Miller [1984] discusses the results of his analysis in the case where the number of jobs in an occupation is infinite (roughly corresponding to infinite branching here). It is unclear whether his results (for finite numbers) extend trivially to the infinite case. That such an extension is non-trivial can be seen in Banks & Sundaram [1992].

[7]In principle, this information could be about the number of branches at descendant nodes, the distribution of returns/costs there, etc. In practice, we will restrict ourselves to the information being solely about *actual* returns at descendant nodes.

in time by a factor β. The agent is facing a finite number of occupations and his problem is to choose a strategy which maximises his expected discounted sum of rewards, net of costs incurred, i.e. he is risk neutral. The maximal expected reward over feasible policies π, denoted by the value function $V(x)$, is given by:

$$V(x(0)) = \sup_{\pi} \mathrm{E}_{\pi}\Big[\sum_{0}^{\infty} \beta^t R(x(t), u(t)) \;\; | \;\; x(0)\Big],$$

where $R(x, u)$ is the immediate reward realised when action u is taken in state x and $x(t)$ is the state of the process at time t, $t = 0, 1, \ldots$.

It was shown in the previous chapter that, under the conditions which obtain here, importantly

- the signal received when a node is explored is informative only about descendant nodes,

- the (bounded) rewards and signals do not depend on the time at which the action is taken,

the agent is facing a bandit super-process, the problem can be reformulated as a dynamic programming one, and an index policy (specifically the Gittins index policy) exists which recommends the optimal action, *despite* the pay-offs to various actions being correlated.

Note that the two conditions essentially guarantee that the state evolves in a Markovian fashion which is stationary or time-homogeneous. Loosely, the first condition (relating to the informativeness of the signals) brings with it a sufficient element of independence, without which the Gittins index policy would not be optimal; the second condition (time-homogeneity) also rules out switching costs, the presence of which would also entail the sub-optimality of the Gittins index policy (see Banks & Sundaram [1994]).

2 Optimal Policy

2.1 Gittins index

Consider just one of the actions available to the agent at time t, and assume there is a fall-back option which has a NPV of m. The agent is facing a stopping problem in which he can either opt for the fall-back (now and forever) or take the available action this period and then once again choose between the fall-back and the action(s)

available at time $t+1$. The smallest m which makes the agent indifferent between stopping and continuation is the *Gittins index* of the action available at time t.

In the current setting, the Gittins index policy is optimal (see Chapter 1), i.e. to each available choice attach an index which depends on the current state of only that choice, then select the choice which currently has the highest index.

Clearly, if a job has a known return stream with NPV w, the Gittins index (reservation value) of that job is simply w. Given a straight choice between two jobs with known return streams, the agent would prefer the job with the higher reward.

2.2 Reservation value – job search

Once the agent has become qualified in a given occupation, he has learnt the component with NPV denoted by y that the returns to him from jobs in this occupation have in common. It is shown in Appendix A that the NPV of the reservation return stream for an arbitrary job in this occupation which costs c_1 to find is given by r_1+y where

$$(1-\beta)r_1 = -c_1 + (1-\beta)\mathrm{E}[z] + \beta\int_{r_1}^{b_1}(1-G_1(z))\,dz$$

and, when you have a fall-back whose NPV is m, the value of having the job opportunity is

$$\phi_1(m,y) = m + \left(0 \vee \left((1-\beta)(r_1+y-m) + \beta\int_{m-y}^{r_1}(1-G_1(z))\,dz\right)\right)$$

(where $m \vee w$ means the larger of m and w). That is, when $m > r_1 + y$ you take the fall-back m; otherwise you pay the cost to find the job and expect a net improvement over the fall-back consisting of two terms: one coming from the return you expect in the first period of employment, and one from the NPV of the larger of the fall-back and the revealed return.[8]

Implications

We can derive some comparative statics results by looking at the above formula for r_1.

- 'Cheaper' is preferred –
  Consider two job opportunities with different search costs, the same mean

[8]The above expression for $\phi_1(m,y)$ simplifies to $m + (0 \vee \int_{m-y}^{r_1} 1-\beta G_1(z)\,dz)$; however, the former expression is more convenient if we wish to make the comparison of choosing among two or more jobs with choosing among two or more occupations.

return, and the same distribution for the return about that mean. Then the agent prefers to try the job with the lower search cost first.

- 'More' is preferred –
  Consider two job opportunities with the same search cost, different mean returns, but the same distribution for the return about that mean. Specifically, $G_1(z) = H_1(z+d)$ for some positive constant d. Then the agent prefers to try the job with the higher expected return first.

- 'Riskier' is preferred –
  Consider two job opportunities with the same search cost, the same mean return, but different CDFs for the return about that mean. Specifically, let H_1 be a mean-preserving spread of G_1 (with the 'single-crossing property'). Then the agent prefers to try the 'riskier' job first.

- 'Patience' is preferred –
  Consider individuals with different discount factors who are facing the same job opportunities. Then the more patient the agent, the higher is his reservation value for each job. However, the preferences of one agent are not necessarily echoed by the others.

The first two results are immediate (and not surprising).

Informally, the third result follows from the following argument: H_1 has more weight in the tails, and so it is lower than G_1 in the upper part of the range of integration. Therefore the integrand involving H_1 (for fixed r_1) is greater, so r_1 must rise to compensate.[9] So, given a straight choice between a risky job and a safe job, the worker tries the risky job first – he is a 'Young Turk' and if there is a bad outcome, he can always try the safe job (and become an 'Old Stager').

As to the fourth result, the derivative of r_1 with respect to β is

$$\frac{\partial r_1}{\partial \beta} = \frac{(r_1 + c_1 - \mathrm{E}[z]) \;/\; \beta}{1 - \beta\, G_1(r_1)} = \frac{\int_{a_1}^{r_1} G_1(z)\, dz}{1 - \beta\, G_1(r_1)} \geq 0\,.$$

However, it is easy to construct examples in which an impatient agent tries a safer job first (rejecting only the worst outcomes), whereas a patient agent tries a riskier

[9] Formally, subtract the expression defining r_G from that defining r_H and rearrange to give

$$(1-\beta)(r_H - r_G) + \beta \int_{r_G}^{r_H} (1 - G_1(z))\; dz = \beta \int_{r_H}^{b_1} (G_1(z) - H_1(z))\; dz$$

The RHS is non-negative for the stated CDFs, implying $r_H \geq r_G$.

job first (accepting only the best outcomes); a third (intermediate) agent will either try the safer job first, but accept only the best outcomes, or try the riskier job first, but reject only the worst outcomes. (See Section 2.4 for another example relating higher discount factors with a preference for riskier opportunities.)

The next two results are for the special case in which all the jobs in the given occupation are *ex ante* identical, i.e. the search cost is the same for each one, and the actual return is drawn from the same distribution.

- Little retraining –
  Once a worker has selected a job in one occupation, he will not switch to a second occupation before having tried all the other jobs in the first occupation.

- Little returning –
  Once a worker has selected a job in one occupation and rejected it, he will not return to that job before having tried all the other jobs in that occupation.

The first result holds because if it is optimal for the worker to select a job in one occupation, then the reservation value for that job must be greater than the reservation value for other occupations and jobs in other occupations. As all the jobs in one occupation are *ex ante* identical, the reservation value for the other jobs in this occupation must also be greater, therefore he would try them first.

The second result holds for a similar reason. If the worker rejects a job after having selected it, it must be because the revealed return is less than the reservation value for the other jobs in this occupation. Therefore he would never return to this job while other jobs in this occupation are untried.

Note that when the number of (*ex ante* identical) job opportunities in the chosen occupation becomes arbitrarily large, the agent will spend at most one spell in any job, and this will last either one period or forever.[10] This is an extreme form of the tenure/ turnover relationship, or duration persistence.

2.3 Reservation value – occupational choice

Now that we have analysed what the agent would do after having selected an occupation, we consider what factors will affect his choice of occupation.

If there are γ identical jobs available in this occupation, then the reservation value (at the job level) is unchanged, but, as is shown in Appendix A, the value of

[10] If the return were not perfectly revealed in the first period of work, then the 'one spell' result would continue to hold, but the 'one period' result would not.

having many opportunities when there is a fall-back whose NPV is M becomes

$$\Phi_1(M, y) = M + \left(0 \vee \int_{M-y}^{r_1} 1 - [\beta G_1(z)]^\gamma \, dz\right).$$

(The expression is much more cumbersome if the jobs are not identical;[11] it was not thought that the analytical complexity of dealing with non-identical jobs would yield additional insights at this level. Thus, for the rest of this section we shall assume that the jobs in any given occupation are *ex ante* identical.)

The NPV of the reservation value stream for embarking on the training for any given occupation can now be calculated. It is r_2, defined implicitly by

$$(1-\beta)r_2 = -c_2 + \beta \int_{r_2-r_1}^{b_2} \left(1 - [\beta G_1(r_2 - y)]^\gamma\right) \left((1 - G_2(y)\right) \, dy$$

and, when you have a fall-back whose NPV is m, the value of having the occupation opportunity is

$$\begin{aligned} \phi_2(m, 0) \;=\; & m + (1-\beta)(r_2 - m) \\ & + \beta \int_{m-r_1}^{r_2-r_1} \left(1 - [\beta G_1(r_2 - y)]^\gamma\right) \left(1 - G_2(y)\right) \, dy \\ & + \beta \int_{m-r_1}^{b_2} \left([\beta G_1(r_2 - y)]^\gamma - [\beta G_1(m - y)]^\gamma\right) \left(1 - G_2(y)\right) \, dy \end{aligned}$$

when $m \leq r_2$ (and you take the opportunity), and m otherwise (when you don't).

Implications

Before the worker has become qualified for an occupation, there is the same preference for lower search costs and higher mean returns, as in the previous subsection, and the same ambiguity regarding the discount factor. Additionally, we have the following results.

- 'More choice' is preferred (comparative statics with respect to γ) –
  Consider two occupations with the same CDF for the common (occupation) component, the same CDF for the particular (job) component, but a different

[11] For example, if there are three jobs with CDFs $F(\cdot)$, $G(\cdot)$ and $H(\cdot)$ and reservation values $r_F < r_G < r_H$, then the above expression for Φ_1 becomes

$$\begin{aligned} \Phi_1(M, y) = M &+ \left(0 \vee \int_{M-y}^{r_H} (1 - \beta H(z)) \, dz\right) \\ &+ \left(0 \vee \int_{M-y}^{r_G} \beta H(z)(1 - \beta G(z)) \, dz\right) + \left(0 \vee \int_{M-y}^{r_F} \beta H(z) \beta G(z)(1 - \beta F(z)) \, dz\right) \end{aligned}$$

and the formula for r_2 becomes correspondingly more complicated.

number of job opportunities. Then the worker trains for the occupation with the greater number of job opportunities first.

- 'Riskier' is preferred (I) (comparative statics with respect to G_1) –
  Consider two occupations with the same number of job opportunities, the same CDF for the common component, but different CDFs for the particular component. Specifically, let H_1 be a mean-preserving spread of G_1 (with the 'single-crossing property'), while $H_2 = G_2$. Then the worker trains for the 'riskier' occupation first.

- 'Riskier' is preferred (II) (comparative statics with respect to G_2) –
  Consider two occupations with the same number of job opportunities, the same CDF for the particular component, but different CDFs for the common component. Specifically, let H_2 be a mean-preserving spread of G_2 (with the 'single-crossing property'), while $H_1 = G_1$. Then the worker trains for the 'riskier' occupation first.

These results follow from looking at the expression for r_2. The first result is quite intuitive and follows from observing that r_2 is increasing in γ.

In the third case, as H_2 has more weight in the tails it is lower than G_2 in the upper part of the range of integration. Therefore the integrand involving H_2 (for fixed r_2) is greater, so r_2 must rise to compensate.

In the second case, as already noted, given a straight choice between a risky job and a safe job, the worker prefers to try the risky job first. The second result then follows from a similar argument to the last, arguing that the integrand *and* the range of integration involving H_1 (for fixed r_2) is greater, so again r_2 must rise to compensate.

Reason: Trying the riskier opportunity first increases the option value of being able to 'back out' of a poor outcome; this follows from the established fact that increased volatility raises option values (see, for example, Dixit & Pindyck [1994]). ♢

2.4 Value of information

In an attempt to determine whether it is simply option values that are driving the agent's apparent preference for trying riskier opportunities first, we look at the following situation.

Consider two occupations: in the first, the common component of the return is very high or very low, but the variance of returns about that common component

is small; in the second, the common component of the return is only quite high or quite low, but the variance of returns about that common component is large. Which occupation should the agent train for first? Does the choice switch as we vary the number of jobs in the occupations? Does the optimal behaviour differ across agents with different discount factors?

Specifically, let us consider two occupations characterised by CDFs G_1 and G_2 on the one hand, and H_1 and H_2 on the other, with a common γ, such that $G_1 = H_2$ (with common search cost) and $G_2 = H_1$ (again with common search cost). Thus, the combined mean and variance is the same in each occupation. In the absence of discounting, and when $\gamma = 1$, we might expect the worker to be indifferent as to which occupation to train for first. If we increase the number of job opportunities, we might then expect the worker to prefer the occupation with the larger variance at the job level. However, in the presence of discounting, the worker may prefer to resolve the major uncertainty first, i.e. prefer the occupation with the *smaller* variance at the job level.

This problem is very difficult to analyse analytically, but we can fix on representative distributions and simulate the outcome numerically. The chosen distributions were as follows and the results we report were obtained under a broad range of parameter values. The safe action was a draw from the discrete distribution on $\{-1, 1\}$ with the outcomes being equally likely; similarly, the risky action was a draw from the discrete distribution on $\{-d, d\}$ for some $d > 1$, specifically 4. A cost of less than $\frac{1}{3}$ was deemed quite low, and a cost of more than $\frac{1}{3}$ was deemed quite high (see below).

In Figure 2.2, the decreasing discount factor is represented on the horizontal axis, and on the vertical axis is the amount by which the reservation value for embarking on 'risky-safe' first exceeds that for embarking on 'safe-risky' first. Thus, there is a clear preference for resolving the major uncertainty first in the following scenario: the costs of each action are the same, and are quite high; there is only one job in each occupation. So, in the absence of discounting, and when $\gamma = 1$, the worker is *not* indifferent as to which occupation to train for first. (However, indifference does obtain in the undiscounted single-job case if the costs remain the same as each other but are quite low.)

Reason: Following either strategy ('safe-risky' or 'risky-safe'), if the agent gets a high realisation at the occupation level, then he will look for a job, but if he gets a low realisation at the occupation level, then he will retrain, in which case the continuation costs and benefits are comparable between the two strategies. In the former case, if he gets a high realisation at the job level as well he will stop; *the important difference arises*

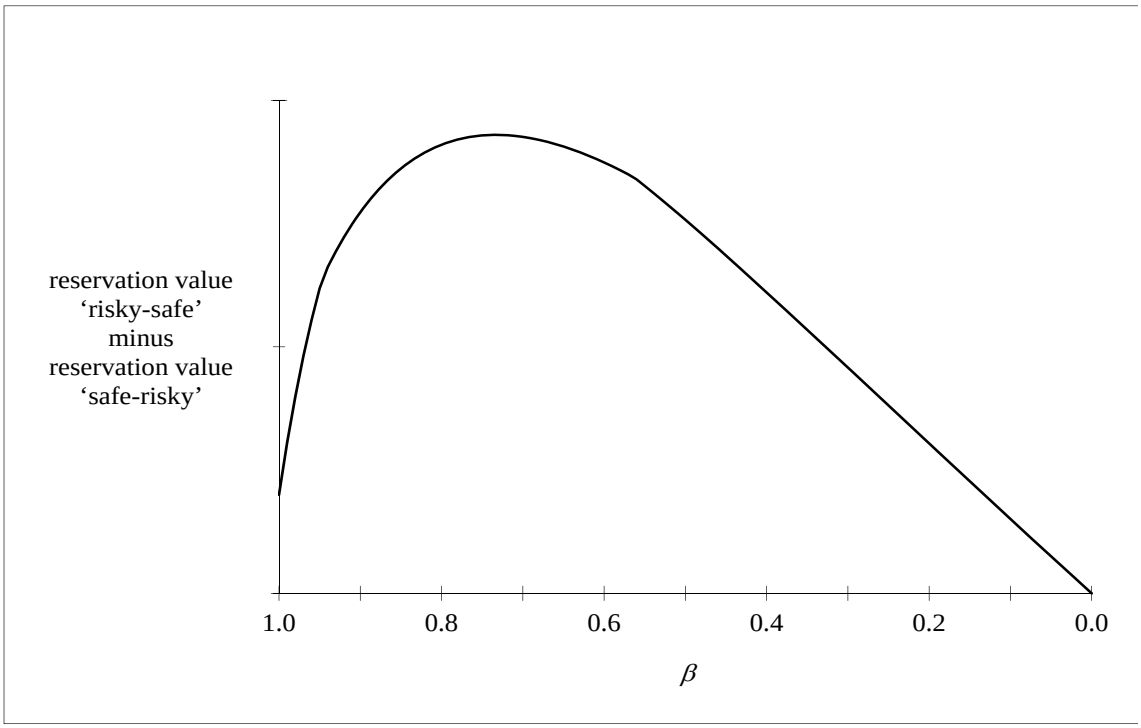


Figure 2.2: Resolve major uncertainty first

if a high realisation at the occupation level is followed by a low realisation at the job level – if he is following a 'safe-risky' strategy, he will train for the other occupation, but if he is following a 'risky-safe' strategy, he will stop if costs are quite high (foregoing a possibly better job in the other untried occupation but incurring no further costs), only training for the other occupation if the costs are quite low. ♢

If we increase the number of job opportunities, the curve does fall in general as we would expect, indicating a weaker preference for embarking on 'risky-safe' first, but remains above the axis unless we increase the cost of the risky action to be greater than that of the safe action, in which case the worker *does* prefer the occupation with the larger variance at the job level *but only* at very high or quite low discount factors.

Reason: Again, the important difference arises if a high realisation at the occupation level is followed by a low realisation at the job level.

When the costs of the safe action are low, he will train for the other occupation under either strategy when β is close to 1: the benefits from either strategy are the same, but the costs of 'risky-safe' are higher.

For intermediate values of β, he will stop sooner if he follows the 'risky-safe' strategy – benefits will be slightly lower than if he were to follow the 'safe-risky' strategy, but costs

are a lot less.

For values of β closer to 0, again he will stop sooner if he follows the 'risky-safe' strategy – benefits will actually be higher than if he were to follow the 'safe-risky' strategy, but this time costs are a lot more. (Remember: the agent trains one period, then finds a job and gets his return the next; when β is small, the training costs dominate.) $\diamondsuit$

So, the suggested conclusion is that in the presence of discounting, the worker does indeed prefer to resolve the major uncertainty first, i.e. prefer the occupation with the *smaller* variance at the job level, for a broad range of discount factors. Having attempted to control for the impact of option values, the agent clearly places a value on information – sooner is better, in general.

2.5 Reservation value – basic education

As indicated in an earlier footnote, we could require that the agent first obtain some basic education and find out his aptitude for work in general. The interested reader is referred to Appendix A for the derivation of the reservation value for such preliminary education.

3 Job Turnover and Expected Returns

A complete characterisation of job turnover and expected returns is a formidable task, depending as it does on the distributions of expected returns in each occupation and the distributions of actual returns in each job. However, we can say something if we focus on the case when the agent has chosen a particular occupation in which to search for a job. Accordingly, for the most part, we address ourselves to the job level and omit the subscript 1.

Assume the agent is in an occupation consisting of γ identical job opportunities, each of which costs c to sample, and each of which results in a return whose NPV is independently drawn from a distribution with support $[a, b]$ and CDF $G(\cdot)$. We want to find the expected number of jobs he samples, the expected return of the job he actually settles on, and also the probability of his finding the best job.

We know that from the previous section that the reservation return stream for each job opportunity (and for remaining in this occupation as a whole) is given by $(1-\beta)r = -c + (1-\beta)\mathrm{E}[z] + \beta \int_r^b (1 - G(z))\, dz$ and that he will stop whenever he finds a job with NPV r or better.

3.1 Job Turnover

Let p be the probability of success with any one job, i.e. $p = 1 - G(r)$, and let $q = 1 - p$, i.e. $q = G(r)$. The probability of stopping after n searches is $q^{n-1}p$ for $n < \gamma$, and the probability of stopping only at the end is $q^{\gamma-1}$. (When there is the possibility of retraining, then it is with probability q^{γ} that the agent will consider switching and compare the best wage he has been offered so far with the reservation value of embarking on a new occupation.) The expected number of searches can then be calculated to be $(1-q^{\gamma})/(1-q)$ or $(1-G(r)^{\gamma})/(1-G(r))$. With arbitrarily many job opportunities, this approaches $1/p$ or $1/(1-G(r))$.

This clearly has implications for job turnover within an occupation. Take, for example, two occupations ('actors' and 'accountants') with the following characteristics:[12] it costs a budding star 15 to find an acting job, and with equal probability of $\frac{1}{3}$ he can become a star and earn 200, play a supporting role and earn 100, or be an extra and receive nothing; or he can find a job as an accountant at a cost of 10, and receive either 170, 150, or 130, again each with probability $\frac{1}{3}$. The reservation value for becoming an actor is 155, and that for an accountant is 145, so, given an agent who is qualified for both occupations, he tries a job in the acting profession first – despite the higher search cost and lower mean return. A simple observation reveals that the probability of failing in any acting job is $\frac{2}{3}$, whereas that the probability of failing in any accounting job is $\frac{1}{3}$. So, if he turns out to be a star, he stays, otherwise he tries other acting jobs. If, say, there are four jobs in each profession, it is not unlikely that he will eventually give up on acting, and settle for accountancy. (There is a small probability that he will do poorly in all the jobs as an accountant and, with slightly different pay-offs, would revert to playing a supporting acting role.) The number of acting jobs he expects to sample is nearly $2\frac{1}{2}$, whereas the number of accounting jobs he expects to sample, conditional on selecting this occupation, is just less than $1\frac{1}{2}$ – the acting profession exhibits much higher turnover.

We shall return to this example in the next subsection, after a word of caution. The derivative of the expected number of searches w.r.t. the probability of failure is

$$\frac{d}{dq}\left(\frac{1-q^{\gamma}}{1-q}\right) > 0.$$

If we were able to say that $H(r_H) > G(r_G)$ whenever $r_H > r_G$ for CDFs characterising the jobs in two different professions, we could then infer that riskier occupations

[12] We consider the undiscounted case for simplicity. This means that the returns are lump-sum as opposed to per-period.

always exhibit greater turnover. However, counter-examples can be constructed: let $G(\cdot)$ be uniformly distributed on the interval $[0, 200]$, and let $H(\cdot)$ be discretely distributed on $\{0, 200\}$ with each outcome having probability $\frac{1}{2}$; assume common costs of 16. Then H is a mean-preserving spread of G and is therefore preferred ($r_H = 168, r_G = 120$); however, $H(r_H) = 0.5 < 0.6 = G(r_G)$, and so the safer occupation exhibits greater turnover.

Note that unemployment is always voluntary in this model; the only time that the agent is not working is when he is training.

3.2 Expected Returns

Let us first consider the undiscounted case.

If we write the probability of stopping only at the end as $q^{\gamma-1}p + q^\gamma$, then we can say that with probability q^γ *all* the jobs were worth less than r and the agent is left with the best of those (if he cannot retrain), and with probability $1 - q^\gamma$ he settled in a job worth r or more. The expected value of the maximum of γ draws, given that they are each less than r is

$$\frac{\int_a^r z\, d(G(z)^\gamma)}{\int_a^r d(G(z)^\gamma)} = r - \frac{\int_a^r G(z)^\gamma\, dz}{G(r)^\gamma}$$

and the expected value of a job worth r or more is

$$\frac{\int_r^b z\, dG(z)}{\int_r^b dG(z)} = r + \frac{\int_r^b (1 - G(z))\, dz}{1 - G(r)}$$

so the expected revenue is

$$q^\gamma \left\{ r - \frac{\int_a^r G(z)^\gamma\, dz}{G(r)^\gamma} \right\} + (1 - q^\gamma) \left\{ r + \frac{\int_r^b (1 - G(z))\, dz}{1 - G(r)} \right\}.$$

Remembering that $q = G(r)$ and $c = \int_r^b (1 - G(z))\, dz$ this can be simplified to

$$r - \int_a^r G(z)^\gamma\, dz + c\,(1 - q^\gamma)/(1 - q).$$

The last term can clearly be seen as the expected search cost, and so the expected net return is given by the first two terms. (This net return might also be obtained from Section 2.3 by simplifying the expression for $\Phi_1(a, 0)$ with $\beta = 1$.)

As the number of opportunities becomes arbitrarily large, you are almost indifferent at the outset between the occupation and a fall-back of r so the expected value

of the occupation approaches r itself. Your expected revenue approaches $r + c/p$, with c/p being the cost of search.

Returning to our example of accountants and actors, we see that the expected net return to an accountant is just less than the reservation value of 145. However, the expected net return to an actor is 120, 135, 143 for $\gamma = 2, 3, 4$. After that it exceeds that of an accountant, showing the value of increased opportunities within an occupation.

However, in the counter-example, as long as $\gamma > 1$, the expected net return in the risky occupation is *always* greater than the reservation value of the safe occupation which is an upper bound on the expected net return there.

The case which incorporates discounting and 'earn-as-you-go' is only a little more complicated. The expected number of searches is unchanged, but the expected search cost becomes $c\,(1 - [\beta q]^\gamma)/(1 - \beta q)$ in the presence of discounting. As noted above, the net return can be obtained from the expression for $\Phi_1(a, 0)$ in Section 2.3. So the expected revenue is

$$r - \int_a^r [\beta G(z)]^\gamma \, dz + c\,(1 - [\beta q]^\gamma)/(1 - \beta q)$$

with the same decomposition as in the undiscounted case.

This device, namely equating the expected net return with the reservation value of the opportunity, can be exploited at other levels, for example when the agent is facing the decision to train for one of two or more occupations, each with its associated range of jobs.

This involves reworking the expression for $\Phi_2(M, y)$ given in Appendix A in a style similar to that offered for $\Phi_1(M, y)$ in the footnote in Section 2.3. For example, assume two occupations characterised by $\mathcal{G} \equiv \langle G_1(\cdot), \gamma_{G_1}, G_2(\cdot)\rangle$ and $\mathcal{H} \equiv \langle H_1(\cdot), \gamma_{H_1}, H_2(\cdot)\rangle$ with $r_{G_2} < r_{H_2}$. (For simplicity, for either occupation, consider its job opportunities to be identical.) Then we have

$$\Phi_2(M, y) = M + \left(0 \vee \int_{M-y}^{r_{H_2}} (1 - \beta J(\mathcal{H}, z))\, dz\right) + \left(0 \vee \int_{M-y}^{r_{G_2}} \beta J(\mathcal{H}, z)(1 - \beta J(\mathcal{G}, z))\, dz\right)$$

where

$$J(\mathcal{F}, z) = 1 - \int_{z - r_{F_1}}^{b_{F_2}} (1 - [\beta F_1(z - z_2)]^{\gamma_{F_1}})\, f_2(z_2)\, dz_2$$

and, writing $a = a_{G_2} \wedge a_{H_2}$, the agent's expected net return is given by $\Phi_2(a, 0)$:

$$r_{H_2} - \int_{r_{G_2}}^{r_{H_2}} \beta J(\mathcal{H}, z)\, dz - \int_a^{r_{G_2}} \beta J(\mathcal{H}, z) \beta J(\mathcal{G}, z)\, dz.$$

3.3 Incomplete Learning

While it is maybe less important to the agent than his expected net return, the probability of finding the best job within an occupation is nevertheless worth a brief investigation.

Complete learning about the chosen occupation, i.e. finding the best job, occurs when either (a) the first draw z is less than r and complete learning about the remaining jobs occurs, or (b) the first draw z is greater than or equal to r and the remaining draws would each be less than or equal to z. So, let $\pi(n)$ be the probability of complete learning when there are n as yet unexplored jobs:[13]

$$\begin{aligned}\pi(n) &= \int_a^r \pi(n-1)\, dG(z) + \int_r^b G(z)^{n-1}\, dG(z) \\ &= \pi(n-1)G(r) + \left(1 - G(r)^n\right)/n\end{aligned}$$

for $n > 1$, with $\pi(1) = 1$. A calculation reveals that $\pi(2) = 1 - \frac{1}{2}\left(1 - G(r)\right)^2$, and so $\pi(1) > \pi(2)$. Observing that

$$\begin{aligned}\pi(n) - \pi(n+1) &= \left(\pi(n-1) - \pi(n)\right) G(r) + \int_r^b \left(G(z)^{n-1} - G(z)^n\right)\, dG(z) \\ &= \left(\pi(n-1) - \pi(n)\right) G(r) + \int_r^b \left(1 - G(z)\right) G(z)^{n-1}\, dG(z)\end{aligned}$$

allows us to conclude, by induction, that $\pi(\cdot)$ is monotonically decreasing.[14] We can now show formally that $\lim_{n\to\infty} \pi(n) = 0$: $\pi(\cdot)$ is bounded below (by 0) and so must have a finite limit which satisfies $\lim_{n\to\infty} \pi(n) = \lim_{n\to\infty} \pi(n-1)G(r)$.

As noted in the introductory paragraph to this section, a similar analysis of Incomplete Learning *across occupations* would be very difficult, and also probably not very fruitful.

Further, it might be interesting to determine the nature of the relationships between Job Turnover, Expected Returns, and Incomplete Learning. However, such an investigation is beyond the scope of this section.

[13] Note that $\pi(1) = 1$ and $\lim_{n\to\infty} \pi(n) = 0$; i.e. if there is only one job, you will take it and it will be the "best", and if there are arbitrarily many jobs you will find one worth at least r in a finite number of searches, and leave a better job untried, almost surely.

[14] In fact, for $n > 1$ and writing q for $G(r)$, it can be shown that

$$\pi(n) = \left(\frac{q^0}{n} + \frac{q^1}{n-1} + \frac{q^2}{n-2} + \ldots + \frac{q^{n-1}}{1}\right) - \left(\frac{1}{n} + \frac{1}{n-1} + \ldots + \frac{1}{2}\right) q^n.$$

4 Conclusion

In this chapter we have developed a model of costly search for an *experience* good, in which returns are correlated *ex ante*. The main results are that newcomers prefer to try riskier ventures first, in which returns are modest and turnover is high, whereas more experienced workers have found a good match – returns are on average greater and turnover is less.

There are two main reasons underlying this behaviour. This first is the *option value* of being able to reverse out of a poor outcome. The second can be seen as a *demand for information* – the resolution of a major source of uncertainty has a higher information content than the resolution of a minor source of uncertainty. These two factors combine to ensure that an occupation with low mean returns might nevertheless be an enticing proposition. If the cost of finding out whether or not he is a 'star' is not too high, then the agent will.

To paraphrase Marshall: "Adventurous young people are more attracted by prospects of success than they are deterred by the fear of failure" – what we have shown here is that this is not necessarily because they are behaving irrationally.

We should mention a recent paper by Prendergast & Stole [1996] with a similar title: "Impetuous Youngsters and Jaded Old-Timers ...". They indeed are putting forward a possible explanation for the sort of individual behaviour discussed here. However, in their model, the agent *knows* his ability and uses his actions to signal this ability to the market; in this chapter, the agent is effectively learning about his ability as he goes on.

Finally, what our agent is learning at the job level is his firm-specific human capital and at the occupation level is something between his general human capital and his firm-specific human capital. (He would learn his general human capital at the 'basic education' level.) Were we to make the firms active players in this matching model (as in Felli & Harris [1996]), then we would be able to address issues such as returns to occupation- and firm-specific human capital, and how the cost of training should be met – by the worker, the firms, or shared. This is a subject for future research.

Appendix

A Derivation of Gittins Indices with Discounting

Throughout this section, m denotes the fall-back, the variable y denotes what the agent has already learnt, and the subscripted variables z denote what the agent is about to learn; $r(y)$ denotes the reservation value given that the agent has learnt y, $r = r(0)$, and $\phi(m, y)$ denotes the value of having a fall-back m and having learnt y. M and Φ correspond to m and ϕ when there are many opportunities stemming from the current option of continuing.

We cover the general case, mentioned in the main text, where the agent must first obtain some basic education, and receives a corresponding return whilst training for an occupation.[15] The Gittins index for the choice at the occupation level which is used in the main text is obtained from the 'Level 2' analysis here by setting $y = 0$ and ignoring the term $(1 - \beta)\mathrm{E}[z_2]$, i.e. he has learnt nothing yet, and does not receive any corresponding return whilst training.

Level 0 – 'stay in a job'

At level 0, the agent has learnt the actual return $(1 - \beta)y$ with NPV y; the (NPV of the) fall-back is m. As usual

$$\phi_0(m, y) = m \vee y$$

and $r_0(y) = r_0 + y$ with $r_0 = 0$.

Level 1 – 'search for a job'

At level 1, the agent can either (a) take the fall-back forever, or (b) pay c_1, get the return today (with expected value $(1-\beta)\mathrm{E}[z_1 + y \mid y]$), reveal z_1, then tomorrow he is at level 0 with $\phi_0(m, z_1 + y)$: he takes the fall-back or the known return forever. So, if he continues:

$$\begin{aligned}
\phi_1(m, y) &= -c_1 + (1 - \beta)\mathrm{E}[z_1 + y \mid y] + \beta\mathrm{E}[\phi_0(m, z_1 + y) \mid y] \\
&= -c_1 + (1 - \beta)\mathrm{E}[z_1] + (1 - \beta)y + \beta \int_{a_1}^{b_1} m \vee (z_1 + y)\, dG_1(z_1) \\
&= -c_1 + (1 - \beta)\mathrm{E}[z_1] + (1 - \beta)y + \beta m + \beta \int_{m-y}^{b_1} (1 - G_1(z_1))\, dz_1
\end{aligned}$$

[15] This implies that $r(y) = r + y$; if we were to include the basic education stage, but exclude the agent's receiving an associated reward whilst training, the index rule would remain optimal but the reservation value $r(y)$ would no longer increase one-for-one with y at 'Level 2'.

From indifference, i.e. $r_1(y) = \phi_1(r_1(y), y)$:

$$\begin{aligned} r_1(y) &= -c_1 + (1-\beta)\mathrm{E}[z_1] + (1-\beta)y + \beta r_1(y) \\ &\quad + \beta \int_{r_1(y)-y}^{b_1} (1 - G_1(z_1))\, dz_1 \\ (1-\beta)(r_1(y) - y) &= -c_1 + (1-\beta)\mathrm{E}[z_1] + \beta \int_{r_1(y)-y}^{b_1} (1 - G_1(z_1))\, dz_1 \end{aligned}$$

An increase in $r_1(y) - y$ makes the LHS increase and the RHS decrease, and the opposite for a decrease in $r_1(y) - y$, so we conclude that $r_1(y) - y$ is constant and $r_1(y) = r_1 + y$ where

$$(1-\beta)r_1 = -c_1 + (1-\beta)\mathrm{E}[z_1] + \beta \int_{r_1}^{b_1} (1 - G_1(z_1))\, dz_1 \tag{A.1}$$

So, when $m > r_1 + y$, you stop with $\phi_1(m, y) = m$, and when $m \leq r_1 + y$, you continue; by subtracting (A.1) from the equation giving $\phi_1(m, y)$ we see that

$$\begin{aligned} \phi_1(m,y) - (1-\beta)r_1 &= (1-\beta)y + \beta m + \beta \int_{m-y}^{r_1} (1 - G_1(z_1))\, dz_1 \\ \phi_1(m,y) &= m + (1-\beta)(r_1 + y - m) + \beta \int_{m-y}^{r_1} (1 - G_1(z_1))\, dz_1 \end{aligned}$$

and in general

$$\begin{aligned} \phi_1(m,y) &= m \vee \left(m + (1-\beta)(r_1 + y - m) + \beta \int_{m-y}^{r_1} (1 - G_1(z_1))\, dz_1 \right) \\ &= m + \left(0 \vee \left((1-\beta)(r_1 + y - m) + \beta \int_{m-y}^{r_1} (1 - G_1(z_1))\, dz_1 \right) \right) \end{aligned} \tag{A.2}$$

Level 2 – 'qualify for an occupation'

In the continuation region for *level 1* ($m \leq r_1 + y$), $\partial\phi_1(m,y)/\partial m = \beta G_1(m - y)$, and in the stopping region $\phi_1(m, y) = m$ as usual, and the partial derivative is 1. Using the formula relating the value of many opportunities to a single opportunity[16]

$$\Phi_1(M, y) = B - \int_M^B \left(\frac{\partial}{\partial m} \phi_1(m, y) \right)^{\gamma_1} dm$$

we have $\Phi_1(M, y) = M$ in the stopping region, and in the continuation region ($M \leq r_1 + y$):

$$\begin{aligned} \Phi_1(M,y) &= B - \int_{r_1+y}^{B} 1^{\gamma_1} dm - \int_M^{r_1+y} [\beta G_1(m-y)]^{\gamma_1}\, dm \\ &= B - \int_{r_1+y}^{B} dm - \int_M^{r_1+y} dm + \int_M^{r_1+y} 1 - [\beta G_1(m-y)]^{\gamma_1}\, dm \\ &= M + \int_{M-y}^{r_1} 1 - [\beta G_1(z)]^{\gamma_1}\, dz \end{aligned}$$

[16] B is the bound on the rewards; γ_1 is the number of opportunities, assumed to be identical. See Theorem 2.1 of the previous chapter.

Thus

$$\Phi_1(M,y) = M + \left(0 \vee \int_{M-y}^{r_1} 1 - [\beta G_1(z)]^{\gamma_1} dz\right) \tag{A.3}$$

So, in the continuation region for level 2:

$$\begin{aligned}
\phi_2(m,y) &= -c_2 + (1-\beta)\mathrm{E}[z_2 + y \mid y] + \beta \mathrm{E}[\Phi_1(m, z_2+y) \mid y] \\
&= -c_2 + (1-\beta)\mathrm{E}[z_2] + (1-\beta)y \\
&\quad + \beta \int_{a_2}^{b_2} m + \left(0 \vee \int_{m-y-z_2}^{r_1} 1 - [\beta G_1(z)]^{\gamma_1}\ dz\right) dG_2(z_2) \\
&= -c_2 + (1-\beta)\mathrm{E}[z_2] + (1-\beta)y \\
&\quad + \beta m + \beta \int_{m-y-r_1}^{b_2} \left(\int_{m-y-z_2}^{r_1} 1 - [\beta G_1(z)]^{\gamma_1}\ dz\right) dG_2(z_2) \\
&= -c_2 + (1-\beta)\mathrm{E}[z_2] + (1-\beta)y + \beta m \\
&\quad + \beta \int_{m-y-r_1}^{b_2} \left(1 - [\beta G_1(m-y-z_2)]^{\gamma_1}\right)\left(1 - G_2(z_2)\right)\, dz_2
\end{aligned}$$

From indifference, i.e. $r_2(y) = \phi_2(r_2(y), y)$:

$$\begin{aligned}
r_2(y) &= -c_2 + (1-\beta)\mathrm{E}[z_2] + (1-\beta)y + \beta r_1(y) \\
&\quad + \beta \int_{r_2(y)-y-r_1}^{b_2} \left(1 - [\beta G_1(r_2(y)-y-z_2)]^{\gamma_1}\right)\left(1 - G_2(z_2)\right)\, dz_2 \\
(1-\beta)(r_2(y)-y) &= -c_2 + (1-\beta)\mathrm{E}[z_2] \\
&\quad + \beta \int_{r_2(y)-y-r_1}^{b_2} \left(1 - [\beta G_1(r_2(y)-y-z_2)]^{\gamma_1}\right)\left(1 - G_2(z_2)\right)\, dz_2
\end{aligned}$$

Arguing that an increase in $r_2(y) - y$ would decrease both the integrand and the range of integration whilst increasing the LHS and vice versa, we see that $r_2(y) - y$ is constant and $r_2(y) = r_2 + y$ where

$$(1-\beta)r_2 = -c_2 + (1-\beta)\mathrm{E}[z_2] + \beta \int_{r_2-r_1}^{b_2} \left(1 - [\beta G_1(r_2-z_2)]^{\gamma_1}\right)\left(1 - G_2(z_2)\right)\, dz_2 \tag{A.4}$$

So, when $m > r_2 + y$, you stop with $\phi_2(m,y) = m$, and when $m \leq r_2 + y$, you continue; by subtracting (A.4) from the equation giving $\phi_2(m,y)$ we see that

$$\begin{aligned}
\phi_2(m,y) &= m + (1-\beta)(r_2 + y - m) \\
&\quad + \beta \int_{m-y-r_1}^{r_2-r_1} \left(1 - [\beta G_1(r_2-z_2)]^{\gamma_1}\right)\left(1 - G_2(z_2)\right)\, dz_2 \\
&\quad + \beta \int_{m-y-r_1}^{b_2} \left([\beta G_1(r_2-z_2)]^{\gamma_1} - [\beta G_1(m-y-z_2)]^{\gamma_1}\right) \\
&\qquad \times \left(1 - G_2(z_2)\right)\, dz_2
\end{aligned} \tag{A.5}$$

Level 3 – 'acquire basic education'

In the continuation region for *level 2* ($m \leq r_2 + y$):

$$\partial\phi_2(m,y)/\partial m = \beta\left(1 - \int_{m-y-r_1}^{b_2} \left(1 - [\beta G_1(m-y-z_2)]^{\gamma_1}\right) g_2(z_2)\, dz_2\right)$$

and in the stopping region $\phi_2(m,y) = m$ as usual, and the partial derivative is 1.

Once again using the formula relating the value of many opportunities to a single opportunity, in the stopping region we have $\Phi_2(M,y) = M$, and in the continuation region ($M \leq r_2 + y$):

$$\begin{aligned}
\Phi_2(M,y) &= B - \int_{r_2+y}^{B} 1^{\gamma_2}\, dm \\
&\quad - \int_{M}^{r_2+y} \left[\beta\left(1 - \int_{m-y-r_1}^{b_2} \left(1 - [\beta G_1(m-y-z_2)]^{\gamma_1}\right) g_2(z_2)\, dz_2\right)\right]^{\gamma_2} dm \\
&= M + \int_{M-y}^{r_2} 1 - \left[\beta\left(1 - \int_{z-r_1}^{b_2} \left(1 - [\beta G_1(z-z_2)]^{\gamma_1}\right) g_2(z_2)\, dz_2\right)\right]^{\gamma_2} dz
\end{aligned}$$

Thus:

$$\Phi_2(M,y) = M + \left(0 \vee \int_{M-y}^{r_2} 1 - \left[\beta\left(1 - \int_{z-r_1}^{b_2} \left(1 - [\beta G_1(z-z_2)]^{\gamma_1}\right) g_2(z_2)\, dz_2\right)\right]^{\gamma_2} dz\right) \tag{A.6}$$

So, in the continuation region for basic education:

$$\phi_3(m,y) = -c_3 + (1-\beta)\mathrm{E}[z_3 + y \mid y] + \beta\mathrm{E}\left[\Phi_2(m, z_3 + y) \mid y\right]$$

$$\begin{aligned}
&\phi_3(m,y) + c_3 - (1-\beta)\mathrm{E}[z_3] - (1-\beta)y - \beta\int_{a_3}^{b_3} m\, dG_3(z_3) \\
&= \beta\int_{a_3}^{b_3} \left(0 \vee \int_{m-y-z_3}^{r_2} 1 - \left[\beta\left(1 - \int_{z-r_1}^{b_2} \left(1 - [\beta G_1(z-z_2)]^{\gamma_1}\right) g_2(z_2)\, dz_2\right)\right]^{\gamma_2} dz\right) \\
&\quad \times dG_3(z_3)
\end{aligned}$$

$$\begin{aligned}
&\phi_3(m,y) + c_3 - (1-\beta)\mathrm{E}[z_3] - (1-\beta)y - \beta m \\
&= \beta\int_{m-y-r_2}^{b_3} \left(\int_{m-y-z_3}^{r_2} 1 - \left[\beta\left(1 - \int_{z-r_1}^{b_2} \left(1 - [\beta G_1(z-z_2)]^{\gamma_1}\right) g_2(z_2)\, dz_2\right)\right]^{\gamma_2} dz\right) \\
&\quad \times dG_3(z_3) \\
&= \beta\int_{m-y-r_2}^{b_3} \left(1 - \left[\beta\left(1 - \int_{m-y-z_3-r_1}^{b_2} \left(1 - [\beta G_1(m-y-z_3-z_2)]^{\gamma_1}\right) g_2(z_2)\, dz_2\right)\right]^{\gamma_2}\right) \\
&\quad \times (1 - G_3(z_3))\, dz_3
\end{aligned}$$

the last line following from integrating by parts as usual. So, from indifference, i.e. $r_3(y) = \phi_3(r_3(y), y)$, we obtain an expression implicitly defining $r_3(y) - y$, the RHS of which is just like the undiscounted case (see Chapter 1, Section 3) except that anything raised to a power γ is first multiplied by β as is the entire integral (cf. the last line above). That being the case, we can argue as before that $r_3(y) - y$ is constant and define r_3 as $r_3(0)$ where:

$$\begin{aligned}
&(1-\beta)r_3 + c_3 - (1-\beta)\mathrm{E}[z_3] \\
&= \beta \int_{r_3-r_2}^{b_3} \left(1 - \left[\beta\left(1 - \int_{r_3-z_3-r_1}^{b_2} \left(1 - [\beta G_1(r_3 - z_3 - z_2)]^{\gamma_1}\right) g_2(z_2)\, dz_2\right)\right]^{\gamma_2}\right) \\
&\quad \times (1 - G_3(z_3))\, dz_3
\end{aligned} \tag{A.7}$$

Similar to above, subtracting (A.7) from the equation giving $\phi_3(m, y)$ now gives:

$$\begin{aligned}
&\phi_3(m,y) = m + (1-\beta)(r_3 + y - m) \\
&+ \beta \int_{m-y-r_2}^{r_3-r_2} \left(1 - \left[\beta\left(1 - \int_{r_3-z_3-r_1}^{b_2} \left(1 - [\beta G_1(r_3 - z_3 - z_2)]^{\gamma_1}\right) g_2(z_2)\, dz_2\right)\right]^{\gamma_2}\right) \\
&\quad \times (1 - G_3(z_3))\, dz_3 \\
&+ \beta \int_{m-y-r_2}^{b_3} \left(\left[\beta\left(1 - \int_{r_3-z_3-r_1}^{b_2} \left(1 - [\beta G_1(r_3 - z_3 - z_2)]^{\gamma_1}\right) g_2(z_2)\, dz_2\right)\right]^{\gamma_2}\right. \\
&\qquad \left. + \left[\beta\left(1 - \int_{m-y-z_3-r_1}^{b_2} \left(1 - [\beta G_1(m - y - z_3 - z_2)]^{\gamma_1}\right) g_2(z_2)\, dz_2\right)\right]^{\gamma_2}\right) \\
&\quad \times (1 - G_3(z_3))\, dz_3
\end{aligned} \tag{A.8}$$

Part II

Optimal Experimentation

Chapter 3

Optimal Experimentation in a Changing Environment*

Introduction

In this chapter, we consider an economic agent whose per-period rewards depend on an unobservable and randomly changing state. Owing to noise, the reward observed after taking an action provides only an imperfect signal of the prevailing state. The agent can improve the information content of this signal by experimenting, that is, by deviating from the myopically optimal action that just maximises current pay-off. When choosing an action, therefore, he has to weigh the long-term informational benefits of experimentation against its short-term opportunity cost.

We are interested in a number of issues. How does the agent's optimal action differ from what is myopically optimal? Is this difference large or small (in a sense to be made precise)? And how well does the agent track the prevailing state? We address these questions in a setting where the agent can finely control the information content of the signals he receives, over a range from zero to some natural upper bound.

Our main result is the identification of two qualitatively very different experimentation regimes. One regime is characterised by large deviations from myopic behaviour, guaranteeing that the signals observed by the agent always contain at least a certain amount of information. This allows him to track the state well, in the sense that his beliefs can come arbitrarily close to the truth. The other regime is characterised by small deviations from myopic behaviour, resulting in signals whose information content can become arbitrarily small. In this regime, the prevailing

*An edited version of this chapter is forthcoming in the *Review of Economic Studies*.

state is tracked poorly: the agent eventually becomes 'trapped' in a strict subset of actions such that, in one of the states, beliefs always stay bounded away from the truth.

Specifically, the agent in our model is a monopolist facing an unknown and changing demand function and maximising expected profits over an infinite horizon. The time parameter is continuous. There are two possible states, each characterised by a linear demand curve, and the transitions between these states are governed by a Markov process. The monopolist knows the slope and intercept of each demand curve and the transition probabilities, but he does not know which demand curve he faces at any given time. At each instant, he chooses from a given interval of feasible quantities, and observes a price which is the 'true' price (derived from the prevailing demand curve) plus noise.[1] Given this noisy signal of the underlying state, the monopolist updates his belief in a Bayesian fashion.

The monopolist can increase the information content of the price signal by moving away from the confounding quantity, that is, the quantity at which the two demand curves intersect; setting the confounding quantity itself leads to a completely uninformative signal. Focusing on the most interesting case, we assume that the confounding quantity lies between the quantities which are myopically optimal in each of the two states. This implies that there is a unique belief – the confounding belief – at which the confounding quantity would be chosen by a myopic agent. The two experimentation regimes are distinguished by the optimal behaviour near this belief.

For a given level of noise, when the discount rate and the intensity of state switching are both low, then experimentation is *extreme*: for beliefs in an interval encompassing the confounding belief, the optimal action is to choose a quantity at the boundary of the feasible set, and the optimal quantity (as a function of the belief) exhibits a jump from one boundary to the other. In this regime, the agent's belief tracks the true state well in the sense explained earlier.

When, for the same level of noise, either the discount rate or the switching intensity is high, then experimentation is *moderate*: the monopolist chooses the confounding quantity at the confounding belief, and quantities relatively close to the myopic ones everywhere else. In this regime, the monopolist eventually becomes trapped into choosing quantities on just one side of the confounding quantity (the

[1] Although we do not pursue such an interpretation in this chapter, we invite the reader to think of the two states as representing fads and fashions, or aggregate income fluctuations that affect the elasticity of demand for the monopolist's output. The noise could represent idiosyncratic taste or income shocks, or small fluctuations in the size of the population served by the monopolist.

side which contains the myopic action corresponding to the long-run average state). In fact, when the monopolist chooses the confounding quantity at the confounding belief, his updating is driven exclusively by the possibility of a change in demand, which pulls his belief in the direction of the long-run average state and prevents it from ever crossing back again. Then, the continually changing state entails his belief sometimes being on the 'wrong' side of the confounding belief, in which case it can never get closer to the true state than the confounding belief – the true state is indeed tracked poorly.

The key to the two regimes is that, of the agent's two conflicting objectives (current reward versus information), one is concave in the choice variable, the other convex. Experimentation is extreme if for some beliefs the combined objective is convex, implying corner solutions – this happens when the frequency of change of the environment and the discount rate are low, so the agent values information highly. When either of these parameters increases, current reward becomes more important; eventually, the combined objective is concave at all beliefs, and we have interior solutions, hence moderate experimentation. At the parameter values where the combined objective just becomes concave throughout, we have a *discontinuous* change in the optimal policy. Thus, a small increase in the variability of the environment can provoke a near cessation of experimentation, with drastic consequences for the process of information acquisition.

If we make the assumption that the confounding quantity does *not* lie between the two myopically optimal quantities, then the direction of widening spreads between the two demand curves is unambiguous, and the monopolist deviates from the myopic action by moving away from the intersection. Experimentation is now moderate for *all* parameter values, and the optimal policy function is continuous and monotonic.

We build upon several strands of the literature on optimal Bayesian learning. A number of authors have identified situations where it is optimal to experiment, and have characterised the agent's strategy as a function of his beliefs. Examples include Prescott (1972), Grossman, Kihlstrom and Mirman (1977) and, more recently, Bertocchi and Spagat (1993), Leach and Madhavan (1993), Mirman, Samuelson and Urbano (1993) and Trefler (1993). These papers do not consider confounding actions, so the different experimentation regimes described here do not arise.

Working in an infinite-horizon setting where the unknown reward function is fixed over time, other authors have focused on the agent's limiting behaviour. The first such model in the economics literature is due to Rothschild (1974), and has

subsequently been extended in a number of different directions; see, for example, McLennan (1984), Easley and Kiefer (1988), Kiefer (1989a), and Aghion, Bolton, Harris and Jullien (1991). A common result of these papers is that the agent's beliefs and actions converge. In the limit, the agent learns everything that is *worth* knowing, so experimentation ceases and no further information is gathered. If there is a confounding action and the agent is impatient, however, beliefs need not converge to a one-point distribution at the true reward function, i.e. learning can remain incomplete. Our moderate experimentation trap extends this incomplete learning result to a changing environment.

Allowing the reward function to change randomly adds more realism in that new data continues to be pertinent, so beliefs continue to evolve, and the agent is not doomed to take the same action for evermore. Moreover, the prior with which the agent starts becomes irrelevant in the long run. Here, we follow Kiefer (1989b), Bala and Kiefer (1990), Balvers and Cosimano (1990, 1993, 1994), Rustichini and Wolinsky (1995) and Nyarko and Olson (1996). However, these authors have either focused on different aspects of the problem, or used frameworks that lent themselves to only limited analytical investigation.

The two papers closest to ours are Kiefer (1989b) and Balvers and Cosimano (1990), both studying a monopolist learning about changing linear demand curves. In a framework with two possible demand curves, Kiefer calculates the value function numerically, illustrates two types of optimal policy (one continuous, one with a jump) and simulates the corresponding sample paths of beliefs and actions. In Balvers and Cosimano's framework, on the other hand, both intercept and slope of the unknown demand curve are given by stochastic processes, so there is in fact a continuum of possible demand curves. This seems more realistic than a two-state model, but the added complexity makes it very hard to obtain analytical results. Moreover, the absence of a confounding action means that their result of sluggish price adjustments has no direct comparison with our main findings.

Rustichini and Wolinsky (1995) use a two-armed bandit framework to study monopoly pricing when the buyers' reservation value changes randomly. Their focus is on non-negligible pricing errors even when the frequency of change is negligible. For certain parameter combinations, learning will cease completely even though the state keeps changing. This can be seen as the analogue in a discrete action space of our moderate experimentation trap.

We depart from the above papers by formulating the problem in continuous time. The advantage of this approach is that it allows us to derive sharp analytical results. We are able to establish key properties of the value function and the optimal policy;

we obtain some analytical comparative statics results; and it is straightforward to characterise the sample path properties of beliefs and optimal actions in each of the two experimentation regimes.[2]

Continuous-time models in the economics literature on Bayesian learning have been pioneered by Smith (1992) and Bolton and Harris (1993). Building on a bandit structure as in Karatzas (1984) and Berry and Fristedt (1985), these authors examine multi-agent learning problems with a fixed distribution of rewards. Smith considers agents that enter sequentially and learn by observing a 'snapshot' of the actions taken by previous generations. He shows that the incomplete learning result going back to Rothschild (1974) is not robust to this form of market learning. While Smith's model does not allow agents to observe each other once they have entered, and thus precludes strategic behaviour, Bolton and Harris focus on the informational externality arising when several agents experiment simultaneously and observe each other's actions and outcomes. Felli and Harris (1996) use a variant of the continuous-time bandit framework to study equilibrium wage dynamics in a setting where two firms and a worker learn about the worker's aptitude to perform firm-specific tasks. We follow these three papers with our specification of Brownian noise and the reliance on the filtering techniques from Liptser and Shiryayev (1977). There are two major differences, however: the problem we study is not of the bandit type, and we allow for a changing environment.[3]

The chapter is organised as follows. After presenting the model in Section 1, we proceed to analyse the monopolist's decision problem as an optimal control problem with his belief as the state variable: we describe the evolution of this belief over time (Section 2), then introduce the corresponding Bellman equation and use it to characterise the value function and optimal quantities (Section 3). The main results of the chapter are in Section 4 where we show that, because there is a confounding belief, the parameter space splits into two regions: one in which experimentation is extreme, the other in which it is moderate. We give a sufficient condition for each regime, and we consider the limiting cases of no state switching and no discounting. Section 5 then briefly discusses the simpler scenario when there is no confounding belief: experimentation is moderate for all parameter values, and the comparative

[2]In particular, it is straightforward to obtain the incomplete learning result from McLennan (1984) in the special case of our model where the state transition rates are zero.

[3]Since the first version of this work was circulated, more authors have adopted the continuous-time setting. Bergemann and Välimäki (1996, 1997) use a bandit framework to study situations where two producers and a continuum of consumers learn about the unknown quality of a new good. Moscarini and Smith (1997) study a single-agent problem of costly sequential experimentation and optimal stopping.

statics results are particularly sharp. A summary and concluding remarks follow in Section 6. Technical results are collected in a series of appendices.

1 The Model

We consider a monopolist producing a non-storable good in continuous time. There are two possible states of demand for this good, $k = 0$ or 1. In state k, the expected per-period demand curve (expected price as a function of quantity) is

$$p = \alpha_k - \beta_k q$$

where α_k and β_k are positive constants. The price which the monopolist actually obtains for his output is the expected price plus some noise term, specified below. The state changes according to a continuous time Markov process with the transition probabilities

$$\Pr(k_{t+\Delta t} = 0 \mid k_t = 0) = 1 - \lambda_0 \Delta t + o(\Delta t), \quad \Pr(k_{t+\Delta t} = 1 \mid k_t = 0) = \lambda_0 \Delta t + o(\Delta t),$$

$$\Pr(k_{t+\Delta t} = 0 \mid k_t = 1) = \lambda_1 \Delta t + o(\Delta t), \quad \Pr(k_{t+\Delta t} = 1 \mid k_t = 1) = 1 - \lambda_1 \Delta t + o(\Delta t)$$

where $\lambda_k \geq 0$ $(k = 0, 1)$. In particular,

$$\Pr\left(k_s = k \ \forall s \in [t, t + \Delta t] \mid k_t = k\right) = \exp(-\lambda_k \Delta t)\,;$$

see Karlin and Taylor (1981, p.146). During production, the monopolist knows the parameters α_k, β_k and λ_k $(k = 0, 1)$, but not the state of demand; furthermore, the noise in realised prices prevents him from directly inferring the true state.

We assume that production has constant marginal cost, normalised to zero without loss of generality, so revenue equals profit. At each time t, the monopolist chooses an output level q_t from an exogenously given interval $Q = [q_{\min}, q_{\max}]$ of feasible quantities.[4] The resulting increment in total revenue is

$$dR_t = q_t \left[(\alpha_{k_t} - \beta_{k_t} q_t)\, dt \ + \ \sigma dZ_t\right]$$

[4]We impose non-negativity constraints on quantities, but not on prices. For reasons of tractability, the noise in realised prices will have full support, so negative prices are possible even if we impose $\bar{q} = \min\{\alpha_0/\beta_0, \alpha_1/\beta_1\}$ as the maximal feasible quantity. However, this is still a natural choice for $q_{\max}$ in many situations. In fact, if $\bar{q} \geq \max\{\alpha_0/2\beta_0, \alpha_1/2\beta_1\}$, we can interpret the expected demand curves in the usual way as meaning $p = \max\{\alpha_k - \beta_k q, 0\}$, and argue that the monopolist will never produce more than $\bar{q}$ since beyond this quantity, the expected current revenue *and* the information content of the price signal decrease.

where Z is a standard Wiener process independent of the process k, and $\sigma > 0$ is a constant known to the monopolist.[5] Thus, $dR_t = q_t\, dP_t$ where dP_t is the increment of a cumulative price process P given by

$$dP_t = (\alpha_{k_t} - \beta_{k_t} q_t)\, dt \;+\; \sigma dZ_t\,.$$

The monopolist derives all his information about the state of demand from observing this price process. Consequently, he is restricted to strategies $\mathbf{q} = \{q_t\}$ such that the action taken at time t depends only on the price history up to that time. The set of all admissible strategies is denoted by $\mathcal{Q}$. (See Appendix A for a formal definition.)

The monopolist's initial belief about the state of demand is characterised by π, his subjective probability that $k_0 = 1$. Given this belief, his objective is to choose $\mathbf{q}$ so as to maximise

$$u^{\mathbf{q}}(\pi) = \mathrm{E}_\pi \left[\int_0^\infty r\, e^{-r\,t}\, dR_t \right]$$

where $r > 0$ is the monopolist's discount rate.[6] Up to the multiplication by r, which expresses the pay-off in per-period terms, $u^{\mathbf{q}}(\pi)$ is the expected present value of the revenue flow from strategy $\mathbf{q}$. Substituting for dR_t, we obtain

$$\begin{aligned} u^{\mathbf{q}}(\pi) &= \mathrm{E}_\pi \left[\int_0^\infty r\, e^{-r\,t}\; q_t\, [(\alpha_{k_t} - \beta_{k_t} q_t)\, dt \;+\; \sigma dZ_t] \right] \\ &= \mathrm{E}_\pi \left[\int_0^\infty r\, e^{-r\,t}\; q_t\, [\alpha_{k_t} - \beta_{k_t} q_t]\, dt \right] \end{aligned}$$

since the stochastic integral with respect to the Wiener process Z has zero expectation.

2 Beliefs

Following a strategy $\mathbf{q} \in \mathcal{Q}$ and observing the associated cumulative price process P, the monopolist updates his beliefs about the state of demand in a Bayesian fashion. Let π_t denote the subjective probability he assigns to state 1 at time t, that is, the conditional probability that $k_t = 1$ given the history of the process P up to t.

By the law of iterated expectations, we have

$$u^{\mathbf{q}}(\pi) = \mathrm{E}_\pi \left[\int_0^\infty r\, e^{-r\,t}\; \mathrm{E}_{\pi_t}[q_t\, (\alpha_{k_t} - \beta_{k_t} q_t)]\, dt \right]$$

[5] This is the continuous-time limit of a revenue equation $\Delta R_t = q_t\, [(\alpha_{k_t} - \beta_{k_t} q_t)\, \Delta t + \sigma\sqrt{\Delta t}\, \varepsilon_t]$ $(t = 0, \Delta t, 2\Delta t, \ldots)$ with $\varepsilon_t \sim \mathrm{IIN}(0, 1)$. A model of this type is examined in Kiefer (1989b).

[6] Later, we also consider the limiting case where $r = 0$.

where

$$\mathrm{E}_{\pi_t}[q_t\,(\alpha_{k_t} - \beta_{k_t} q_t)] = q_t\,[(1-\pi_t)\alpha_0 + \pi_t\alpha_1 - ((1-\pi_t)\beta_0 + \pi_t\beta_1)q_t]$$

is the expected revenue, given the observed price history, for quantity q_t. To simplify the notation, we introduce the functions

$$\begin{aligned}\alpha(\pi) &= (1-\pi)\alpha_0 + \pi\alpha_1,\\ \beta(\pi) &= (1-\pi)\beta_0 + \pi\beta_1\end{aligned}$$

which describe the expected intercept and slope of the demand curve given the belief π, and

$$R(\pi, q) = q\,[\alpha(\pi) - \beta(\pi)q]$$

which is the corresponding expected revenue from setting quantity q. Thus, we have the representation

$$u^{\mathbf{q}}(\pi) = \mathrm{E}_\pi\left[\int_0^\infty r\,e^{-r\,t}\;R(\pi_t, q_t)\,dt\right] \tag{1}$$

which does not involve the stochastic variable k_t any more; instead, expected total pay-off is described as a function of beliefs alone.

This suggests looking at strategies based exclusively on the information contained in beliefs. In the next section we show that optimal strategies are in fact stationary Markov strategies, namely ones where the quantity chosen at time t is a (time-invariant) function of the belief at that time, that is, $q_t = q(\pi_t)$. However, first we have to investigate how beliefs evolve over time.

To this end, we define

$$\lambda(\pi) = (1-\pi)\lambda_0 - \pi\lambda_1$$

and

$$\Sigma(\pi, q) = \sigma^{-1}\pi(1-\pi)(\Delta\alpha - \Delta\beta\,q)$$

where $\Delta\alpha = \alpha_1 - \alpha_0$ is the difference in intercepts and $\Delta\beta = \beta_1 - \beta_0$ the difference in slopes between the two expected demand curves. Then, it follows from Liptser and Shiryayev (1977, Chapter 9) that given a strategy $\mathbf{q} \in \mathcal{Q}$, the beliefs evolve according to the filtering equation

$$d\pi_t = \lambda(\pi_t)\,dt + \Sigma(\pi_t, q_t)\,dZ_t^{\mathbf{q}} \tag{2}$$

where $dZ_t^{\mathbf{q}}$ is the increment of a Wiener process. In other words, the change in

beliefs $d\pi_t$ is normally distributed with mean $\lambda(\pi_t)\,dt$ and variance $\Sigma^2(\pi_t, q_t)\,dt$.

Equation (2) emphasises the two separate forces which drive the updating. The drift term $\lambda(\pi_t)\,dt$ takes account of the possibility that the state may change over the next infinitesimal period of time. Given the current belief π, the monopolist assigns probability $1-\pi$ to state 0, hence probability $(1-\pi)\lambda_0$ to a transition from state 0 to state 1 over the next instant dt; in the same way, he assigns probability $\pi\lambda_1$ to a transition from state 1 to state 0. The first possibility increases the probability of being in state 1 after the time dt has elapsed, the second reduces it, and the combined effect leads to the drift term in (2). If at least one of the transition intensities λ_0, λ_1 is nonzero, the linear function λ is downward sloping and vanishes at the *invariant belief*

$$\tilde{\pi} = \frac{\lambda_0}{\lambda_0 + \lambda_1}\,.$$

In view of this, we let $\Lambda = \lambda_0 + \lambda_1$ and rewrite this function as

$$\lambda(\pi) = -\Lambda\,(\pi - \tilde{\pi})\,.$$

This representation shows that state switching introduces mean reversion into the evolution of beliefs. Throughout the chapter, we shall fix an invariant belief $\tilde{\pi}$ and use the parameter Λ to measure the intensity of demand curve switches, and hence the instability of the environment in which the monopolist operates.

The diffusion term $\Sigma(\pi_t, q_t)\,dZ_t^{\mathbf{q}}$ captures the influence of the observed price signal on the evolution of beliefs. $Z^{\mathbf{q}}$ being a Wiener process, this part of the updating is completely unpredictable. Intuitively, this expresses the fact that the current belief already incorporates everything that there is to know, so any change must come as a surprise. The representation

$$dZ_t^{\mathbf{q}} = \sigma^{-1}\left((\alpha_{k_t} - \beta_{k_t} q_t)\,dt \;+\; \sigma dZ_t \;-\; [\alpha(\pi_t) - \beta(\pi_t) q_t]\,dt\right) \tag{3}$$

from Liptser and Shiryayev (1977, Chapter 9) confirms this, showing that the change in beliefs depends on the difference between the realised price, $(\alpha_{k_t} - \beta_{k_t} q_t)\,dt + \sigma dZ_t$, and the expected price, $[\alpha(\pi_t) - \beta(\pi_t) q_t]\,dt$. The greater the spread $\Delta\alpha - \Delta\beta\, q_t$ between the two demand curves, and the lower the noise level σ, the more informative is the price signal, and the more pronounced is the change of beliefs after the signal is observed. There is the possibility of a completely uninformative signal if $\hat{q} = \Delta\alpha/\Delta\beta$ is a feasible quantity – the expected price for this quantity is always the same, regardless of the current state of demand or the current belief. Accordingly, $\Sigma(\pi, \hat{q}) = 0$ for all π. On the other hand, for $\pi = 0$ or 1 the agent is subjectively

certain of the current state and ignores the price signal: $\Sigma(0,q) = 0$ and $\Sigma(1,q) = 0$ no matter which action is taken.

Finally, note that we can simplify the expression on the right-hand side of (3) to $\sigma^{-1}(k_t - \pi_t)(\Delta\alpha - \Delta\beta\, q_t)\, dt \;+\; dZ_t$ and use this to replace $dZ_t^{\mathbf{q}}$ in (2):

$$d\pi_t = \Big\{\lambda(\pi_t) + \sigma^{-2}\pi_t(1-\pi_t)(k_t-\pi_t)(\Delta\alpha - \Delta\beta\, q_t)^2\Big\}\, dt + \Sigma(\pi_t, q_t)\, dZ_t\,. \qquad (4)$$

Looking at the term which contains the factor $k_t - \pi_t$, we see that whenever the signal is informative and the agent is not already subjectively certain, his belief is pulled towards the truth.

3 The Bellman Equation

The representation (1) for the pay-off $u^{\mathbf{q}}(\pi)$, the filtering equation (2) for the evolution of beliefs and the fact that $Z^{\mathbf{q}}$ is a Wiener process allow us to consider the monopolist's decision problem as a problem of optimal control of a diffusion process, the diffusion in question being the process of beliefs. Following the standard approach to this type of control problem,[7] we now turn to the corresponding value function and Bellman equation.

As usual, the value function is defined as

$$u^*(\pi) = \sup_{\mathbf{q}\in\mathcal{Q}}\; u^{\mathbf{q}}(\pi) \qquad (5)$$

for $\pi \in [0,1]$. It is clearly bounded and, being the upper envelope of linear pay-off functions $u^{\mathbf{q}}$, it is also continuous and convex, convexity expressing the fact that information is valuable to the agent.[8] (See Appendix B for details.)

Standard results imply that the value function has further regularity properties, principally that it has a continuous first derivative on $[0,1]$, and a non-negative locally bounded second derivative almost everywhere on $]0,1[$. Moreover, u^* is a solution of the Bellman equation

$$\max_{q\in Q}\Big\{\tfrac{1}{2}\,\Sigma^2(\pi,q)\,u''(\pi) + \;\lambda(\pi)\,u'(\pi) - r\,u(\pi) + r\,R(\pi,q)\Big\} = 0 \qquad (6)$$

[7] See for instance Fleming and Rishel (1975) and Krylov (1980).

[8] Suppose that new information could shift the monopolist's prior to π_1 with probability η and to π_2 with probability $1-\eta$. Then he obtains the expected pay-off $u^*(\eta\pi_1+(1-\eta)\pi_2)$ if he must choose a strategy without the new information, while he can expect the pay-off $\eta u^*(\pi_1) + (1-\eta)u^*(\pi_2)$ if he is allowed to choose after the information is revealed. The latter dominates the former if and only if u^* is convex.

almost everywhere; see Appendix C for details.

We give a brief, heuristic derivation of the Bellman equation. From the Principle of Optimality, we see that u^* satisfies

$$u(\pi) = \max_{q \in Q} \left\{ r\, R(\pi, q)\, dt \;+\; e^{-r\, dt}\; \mathrm{E}_\pi \left[u(\pi + d\pi) \right] \right\} \tag{7}$$

where the first term is the expected current pay-off, and the second term is the discounted expected continuation value. With regard to the latter, we can approximate $e^{-r\, dt}$ by $1 - r\, dt$, and, when u is sufficiently differentiable, Itô's lemma gives us

$$\mathrm{E}_\pi[u(\pi + d\pi)] = u(\pi) + u'(\pi)\, \mathrm{E}_\pi[d\pi] + \tfrac{1}{2}\, u''(\pi)\, \mathrm{E}_\pi[(d\pi)^2]\,.$$

From equation (2), we see that $\mathrm{E}_\pi[d\pi] = \lambda(\pi)\, dt$ and $\mathrm{E}_\pi[(d\pi)^2] = \Sigma^2(\pi, q)\, dt$. The discounted expected continuation value is therefore

$$(1 - r\, dt) \left(u(\pi) + \lambda(\pi)\, u'(\pi)\, dt + \tfrac{1}{2}\, \Sigma^2(\pi, q)\, u''(\pi)\, dt \right) .$$

Substituting this into (7) and ignoring terms of order $(dt)^2$, we obtain

$$u(\pi) = \max_{q \in Q} \left\{ r\, R(\pi, q)\, dt + u(\pi) - r\, u(\pi)\, dt + \lambda(\pi)\, u'(\pi)\, dt + \tfrac{1}{2}\, \Sigma^2(\pi, q)\, u''(\pi)\, dt \right\}$$

which, after simplifying, yields the Bellman equation (6).

The Bellman equation is our main tool for constructing optimal strategies which will in fact be stationary Markov strategies. Such a Markov strategy is derived from a policy function $q : [0, 1] \to Q$ by selecting the quantity $q_t = q(\pi_t)$ when π_t is the belief at time t. A policy function is admissible if this procedure leads to an admissible strategy $\mathbf{q} \in \mathcal{Q}$ for any given initial belief π_0; in Appendix A, we present some regularity conditions which ensure that a given policy function is admissible, principally that either it is Lipschitz continuous, or that it is measurable and the information content of the price signal, as given by $\Delta\alpha - \Delta\beta\, q(\pi)$, is bounded away from zero.

We can now state the following version of the standard *verification theorem*. Suppose that $u : [0, 1] \to I\!R$ solves the Bellman equation (subject to boundary conditions which we shall discuss in the next subsection); suppose further that $q : [0, 1] \to Q$ is an admissible policy function such that

$$q(\pi) \in \arg\max_{q \in Q} \left\{ \tfrac{1}{2}\, \Sigma^2(\pi, q)\, u''(\pi) + \;\lambda(\pi)\, u'(\pi) - r\, u(\pi) + r\, R(\pi, q) \right\}$$

for all π. Then u is the value function and q an optimal policy; see Appendix C.

Next, we discuss the economics behind the Bellman equation.

3.1 Information and Experimentation

Some economic insights can be gained from rewriting the Bellman equation as

$$u(\pi) = \lambda(\pi)\frac{u'(\pi)}{r} + \max_{q \in Q}\left\{\Sigma^2(\pi, q)\frac{u''(\pi)}{2r} \; + \; R(\pi, q)\right\} \tag{8}$$

where the maximisation problem immediately indicates the fundamental trade-off between information gathering and myopic profit maximisation. We look at the three terms on the right-hand side of (8) in turn.

The first term, $\lambda(\pi)u'(\pi)/r$, represents the contribution purely owing to state switching. According to (2), $\lambda(\pi)$ indicates the magnitude and direction of the likely change in belief due to possible state switching, and this (passively acquired) element has the shadow price $u'(\pi)/r$. The resulting contribution to the value function is positive if the belief is expected to move in the direction which increases value.

The next term, $\Sigma^2(\pi, q)\, u''(\pi)/2r$, represents the value of information actively acquired by the agent. Indeed, the discussion after equation (3) above shows that $\Sigma^2(\pi, q)$ provides a measure of the informativeness of the price signal, taking into account the precision of the current belief. This informativeness is valued with the shadow price $u''(\pi)/2r$. Note that for $\Delta\beta \neq 0$ and $u''(\pi) > 0$, the value of information at $\pi \neq 0,\ 1$ is a strictly convex quadratic in q with a global minimum of 0 at $\hat{q} = \Delta\alpha/\Delta\beta$. In particular, it increases strictly with the distance between q and $\hat{q}$.

The last term, $R(\pi, q)$, represents the myopic pay-off. Note that $R(\pi, q)$ is a strictly concave quadratic in q with a global maximum of

$$m(\pi) = \max_q R(\pi, q) = \frac{\alpha(\pi)^2}{4\beta(\pi)}$$

at the quantity

$$q^m(\pi) = \arg\max_q R(\pi, q) = \frac{\alpha(\pi)}{2\beta(\pi)}.$$

We call the functions m and q^m the *myopic optimum pay-off* and the *myopic policy function*, respectively. As

$$R(\pi, q) = m(\pi) - \beta(\pi)[q - q^m(\pi)]^2\,, \tag{9}$$

the myopic pay-off decreases strictly as the distance between q and $q^m(\pi)$ increases.

So, the agent's problem is to choose a quantity that maximises the sum of the value of information actively acquired and the myopic pay-off. This sum is also a quadratic in q and its convexity/concavity depends on whether or not the convexity of the value of information term dominates the concavity of the myopic pay-off. If it does, then an extreme quantity ($q_{\max}$ or $q_{\min}$) will be chosen; otherwise, the optimal choice can be an interior solution. This is the key to the discontinuity in optimal behaviour which we will find in Section 4.

Throughout the chapter, we assume that Q^m, the range of the myopic policy function q^m, is contained in Q, so the myopically optimal quantity is always in the monopolist's choice set. Evaluating the maximand in (8) for the value function $u*$ at the quantity $q = q^m(\pi)$, we thus obtain the inequality

$$u^*(\pi) - \lambda(\pi)\frac{(u^*)'(\pi)}{r} \geq m(\pi)\,. \tag{10}$$

The monopolist is said to *experiment* at the belief π if he deviates from the quantity $q^m(\pi)$. This might render the price signal more informative, but it entails an opportunity cost as is evident from (9). The Bellman equation shows that such a deviation is profitable at a belief π if and only if the inequality (10) is strict at that belief.

Economic intuition suggests that the monopolist will not experiment when he is subjectively certain of the current state of demand. For π tending to 0 or 1, we therefore expect that the value of information $\Sigma^2(\pi, q)\,(u^*)''(\pi)/2r$ tends to zero for all possible quantities q. If this is the case, then the myopic quantity $q^m(0)$ or $q^m(1)$ will be optimal in (8) for $\pi = 0$ or 1, respectively, and formally taking limits in (8), we obtain the following boundary conditions for the value function:

$$u^*(0) - \lambda(0)\frac{(u^*)'(0)}{r} = m(0)\,, \qquad u^*(1) - \lambda(1)\frac{(u^*)'(1)}{r} = m(1)\,.$$

This intuition is confirmed in Appendix C where we show that these boundary conditions are indeed satisfied by the value function. Hence the agent does not experiment at the beliefs 0 and 1. We will see that there is at most one non-degenerate belief at which the inequality (10) might fail to be strict; this belief is identified next.

3.2 The Confounding Quantity and Belief

The quantity at which the demand curves intersect is $\Delta\alpha/\Delta\beta$, denoted by $\hat{q}$, and the corresponding price is $(\alpha_0\beta_1 - \alpha_1\beta_0)/\Delta\beta$, denoted by $\hat{p}$. Unless stated otherwise,

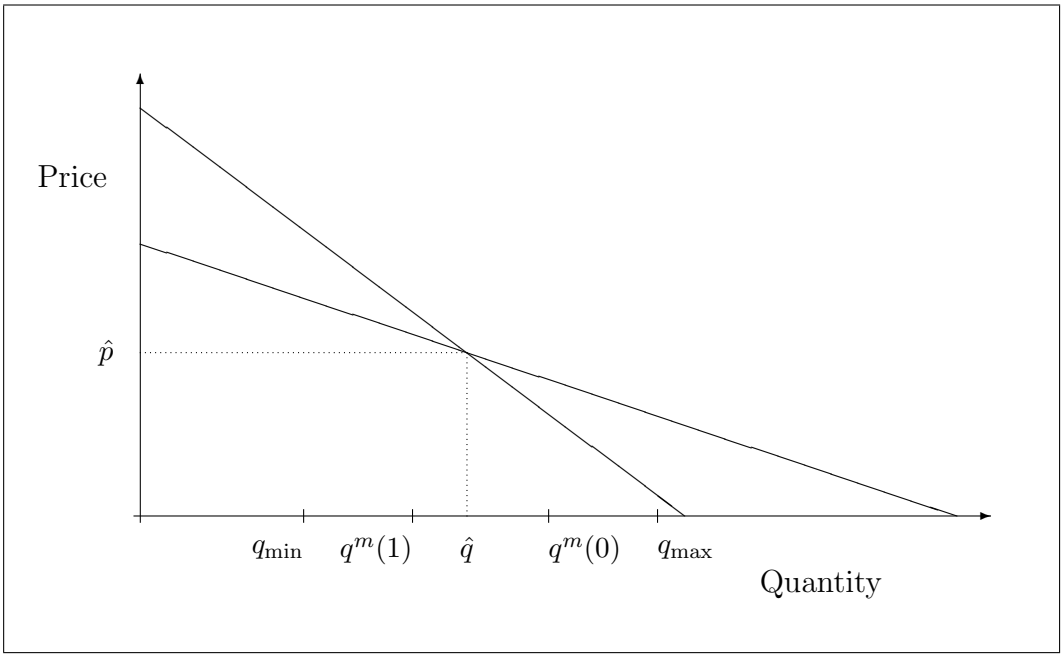


Figure 3.1: The two demand curves

we will make the following

Assumption *The quantity $\hat{q}$ lies strictly between $q^m(0)$ and $q^m(1)$.*

To be more concrete, we will assume without loss of generality that the demand curve in state 1 is steeper than the demand curve in state 0, that is, $\Delta\beta > 0$. With this convention, the assumption amounts to the inequalities $\Delta\alpha > 0$ and $q^m(0) > \hat{q} > q^m(1)$, which implies that $\hat{p} \neq 0$; see Figure 3.1.

We saw in Section 2 that choosing the quantity $\hat{q}$ leads to a completely uninformative price signal – the expected price for this quantity is $\hat{p}$ regardless of the state of demand or the current belief. As this constitutes a confounding action in the sense of Easley and Kiefer (1988), we shall refer to $\hat{q}$ as the *confounding quantity*.

If the monopolist were to choose $\hat{q}$, then he would acquire no information, and so for this action to be optimal it must maximise his myopic pay-off; that is, given the belief π, $\hat{q}$ can be optimal only if $\hat{q} = q^m(\pi)$. Straightforward algebra shows strict monotonicity of the myopic policy function,[9] so there is a unique belief, denoted by $\hat{\pi}$ and called the *confounding belief*, such that $q^m(\hat{\pi}) = \hat{q}$.[10]

For future reference, we define $\hat{m} = \hat{p}\hat{q}$. Clearly, $\hat{m} = R(\pi, \hat{q})$, so we can interpret

[9]Differentiating q^m leads to $(q^m)'(\pi) = -\frac{1}{2}\,\hat{p}\,\Delta\beta/\beta(\pi)^2$ and hence q^m is strictly monotonic whenever $\hat{p} \neq 0$.

[10]A simple calculation reveals that $\hat{\pi} = \alpha_0/\Delta\alpha - 2\,\beta_0/\Delta\beta$.

it as the expected revenue, given *any* belief π, from choosing the quantity $\hat{q}$. In particular, $\hat{m} = m(\hat{\pi})$, and it is easy to verify that this is the global minimum of the myopic pay-off function m; in fact, m is strictly decreasing on $[0, \hat{\pi}]$ and strictly increasing on $[\hat{\pi}, 1]$.

We noted in the previous subsection that for any current belief π, the agent maximises the sum of two quadratics, one convex centred on $\hat{q}$ and one concave centred on $q^m(\pi)$. At the belief $\hat{\pi}$, these quadratics are thus centred on the same quantity, as is their sum. Therefore, at the confounding belief we expect either extreme experimentation (as usual when the combined quadratic is convex) or *no* experimentation (when the combined quadratic is concave). As noted above, which of these two possibilities arises will depend on whether (10) is a strict inequality at $\hat{\pi}$ (extreme experimentation), or holds with equality at $\hat{\pi}$ (no experimentation).

Note that under our assumption, the situation faced by the monopolist satisfies the two *necessary* conditions for experimentation identified by Mirman, Samuelson and Urbano (1993) in a two-period framework: experimentation is informative since a change in quantity affects the informativeness of the price signal ($\Delta\beta \neq 0$); and information is valuable in the sense that different quantities are optimal in the two states ($q^m(0) \neq q^m(1)$). We briefly look at the two special cases of our model where one or other of the conditions is violated.

The case of uninformative experimentation. If we suppose that the two demand curves have the same slope parameter, $\beta_0 = \beta_1$, then the monopolist is facing an unknown and possibly changing intercept.[11] As the demand curves are parallel, a change in output does not affect the spread between the two possible price distributions, so the quantity choice has no impact on the informativeness of the price signal. This renders experimentation uninformative, so the agent has no incentive to deviate from the myopic optimum. Indeed, for $\Delta\beta = 0$, equation (8) reduces to

$$u(\pi) = \lambda(\pi)\frac{u'(\pi)}{r} + \left(\sigma^{-1}\pi(1-\pi)\,\Delta\alpha\right)^2 \frac{u''(\pi)}{2r} + \max_{q\in Q} R(\pi, q),$$

implying that $q^m(\pi)$ is optimal for all π.

The case of worthless information. If we suppose that one and the same quantity is optimal under either demand curve, then $q^m(0) = q^m(1)$, which we denote by $q^\ddagger$. (This happens if and only if the two demand curves intersect exactly on the quantity axis, that is, $\hat{p} = 0$ or, equivalently, $\alpha_0/\beta_0 = \alpha_1/\beta_1$.) In this situation,

[11] Of course, we assume $\alpha_0 \neq \alpha_1$ to avoid trivialities.

information is clearly worthless, and we expect the optimal policy to be $q(\pi) = q^\ddagger$, which is constant over time and across beliefs. To verify this, note that the myopic optimum pay-off function m is linear in this case, so we can find a linear solution u to the Bellman equation.[12] As $u''(\pi) = 0$ throughout, there is no value to information, hence again no incentive to experiment.

In fact, the two conditions that experimentation is informative ($\Delta\beta \neq 0$) and that information is valuable ($q^m(0) \neq q^m(1)$) are *sufficient* for experimentation to occur at almost all beliefs π. To see this, note that $q^m(0) \neq q^m(1)$ implies strict convexity of the myopic pay-off function m, which in turn implies strict convexity of the value function u^*.[13] In particular, we have $(u^*)'' > 0$ almost everywhere. As $\Delta\beta \neq 0$, this means that the myopic quantity $q^m(\pi)$ violates the first order condition for the maximisation problem in (8) at almost all π. At the same time, this shows that the inequality (10) is strict almost everywhere.

3.3 A Differential Equation for the Value Function

The next step in solving the agent's problem is to use the Bellman equation to derive an ordinary differential equation for the value function. The obvious way to do this is in two stages: first calculate optimal quantities in terms of π, $u(\pi)$, $u'(\pi)$ and $u''(\pi)$; then insert these back into the Bellman equation and solve for $u''(\pi)$. However, starting with a simple reformulation of the Bellman equation enables us to get to the desired ODE more directly.

Introducing the notation

$$\tau(\pi) = \Delta\beta^2\, \sigma^{-2}\, \pi^2(1-\pi)^2$$

so that $\Sigma^2(\pi, q) = \tau(\pi)\,[q - \hat{q}]^2$, we can rewrite (8) as

$$u(\pi) - \lambda(\pi)\frac{u'(\pi)}{r} = \max_{q \in Q} \left\{ \tau(\pi)\,[q-\hat{q}]^2\, \frac{u''(\pi)}{2r} + R(\pi, q) \right\}. \tag{11}$$

As $R(\pi, \hat{q}) = \hat{m}$, $\hat{q}$ is suboptimal in (11) as long as $u(\pi) - \lambda(\pi)u'(\pi)/r > \hat{m}$. Under

[12] In fact, m itself solves the Bellman equation if $\Lambda = 0$.

[13] If we assume that u^* is not *strictly* convex, i.e., there is some interval where it is linear, then, in this interval, $(u^*)'$ is constant and $(u^*)''$ is 0. Referring to the maximisation problem in (8), we see that on the one hand, $(u^*)''(\pi) = 0$ implies that the maximum is $m(\pi)$ for all π in this interval, while on the other hand, $(u^*)'$ being constant implies that both the other two terms are linear, which contradicts the strict convexity of m.

this condition, (11) is then easily seen to be equivalent to

$$\tau(\pi)\,\frac{u''(\pi)}{2r} = \min_{q\in Q-\{\hat{q}\}} \frac{u(\pi)-\lambda(\pi)u'(\pi)/r - R(\pi,q)}{[q-\hat{q}]^2}\,, \tag{12}$$

and a quantity $q^* \in Q-\{\hat{q}\}$ attains the maximum in (11) if and only if it attains the minimum in (12).[14]

With v being a generic variable standing for $u(\pi)-\lambda(\pi)u'(\pi)/r$, this observation effectively reduces the analysis of the Bellman equation to the analysis of the function

$$G(\pi,v) = \min_{q\in Q-\{\hat{q}\}} \frac{v-R(\pi,q)}{[q-\hat{q}]^2} \tag{13}$$

and the correspondence

$$O(\pi,v) = \arg\min_{q\in Q-\{\hat{q}\}} \frac{v-R(\pi,q)}{[q-\hat{q}]^2}$$

for (π,v) lying in the set

$$\mathcal{A} = \{(\pi,v)\in\,]0,1[\,\times I\!R:\ v\geq m(\pi) \text{ and } v>\hat{m}\}\,.$$

Note that the condition $v>\hat{m}$ rules out exactly the point $(\hat{\pi},\hat{m})$, i.e. the lowest point on the graph of the myopic pay-off function m; see Figure 3.2.

The function G is well-defined on $\mathcal{A}$, that is, O is nonempty-valued, and we shall see below that G is continuous on $\mathcal{A}$.[15] This implies that the value function u^* is twice differentiable and solves the ODE

$$\tau(\pi)\frac{u''(\pi)}{2r} = G\left(\pi,\ u(\pi)-\lambda(\pi)\frac{u'(\pi)}{r}\right) \tag{14}$$

at least on $]0,1[\,-\{\hat{\pi}\}$, and on the whole of $]0,1[$ if $u^*(\hat{\pi})-\lambda(\hat{\pi})(u^*)'(\hat{\pi})/r>\hat{m}$.[16]

Conversely, we can rephrase the verification theorem as follows. Suppose the function u has a continuous first derivative on $[0,1]$ and solves (14) on $]0,1[\,-\{\hat{\pi}\}$ with the boundary conditions $u(0)-\lambda(0)u'(0)/r=m(0)$ and $u(1)-\lambda(1)u'(1)/r=$

[14] A detailed derivation of this equivalence is given in Appendix D.

[15] We shall give an explicit expression for G which makes continuity obvious. Alternatively, we could show continuity by applying standard arguments which are used in the proof of Berge's Maximum Theorem.

[16] As $G(\pi,\ u^*(\pi)-\lambda(\pi)(u^*)'(\pi)/r$ is continuous in π as long as $u^*(\pi)-\lambda(\pi)(u^*)'(\pi)/r>\hat{m}$, this statement follows directly from Corollary C.1 in the Appendix.

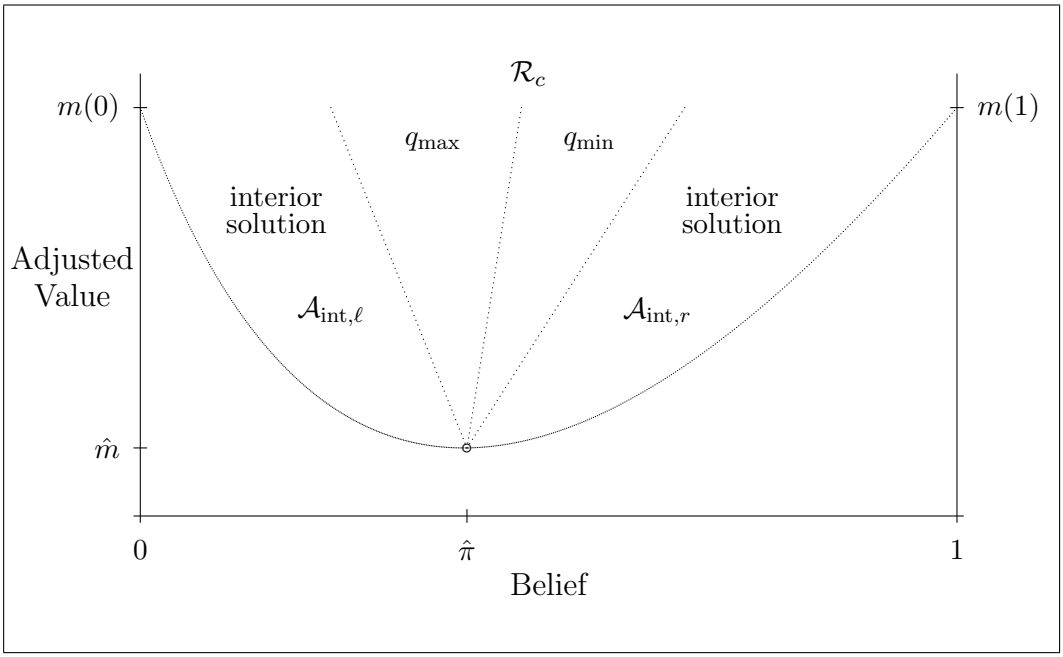


Figure 3.2: The four regions

The convex curve is the myopic pay-off $v = m(\pi)$.

$m(1)$; moreover, suppose that there is an admissible policy function q^* such that

$$q^*(\pi) \in O\left(\pi,\ u(\pi) - \lambda(\pi)\frac{u'(\pi)}{r}\right)$$

for all $\pi \neq \hat{\pi}$. Then $u = u^*$, and the policy function q^* is optimal. (This follows directly from Proposition C.2 in the Appendix.)

3.4 Optimal Quantities

We turn now to a more explicit analysis of the function G and the optimal quantity correspondence O. We just outline the general structure; details are given in Appendix D. The area $\mathcal{A}$ can be divided into four regions by rays emanating from $(\hat{\pi}, \hat{m})$, as in Figure 3.2. The regions which border on the curve $v = m(\pi)$ are associated with the minimisation problem in (13) having an interior solution, and the other two are associated with it having a corner solution. In brief, moving clockwise from the left, we shall have: interior solution, corner solution $q_{\max}$, corner solution $q_{\min}$, interior solution.

The leftmost ray which separates the first two regions goes up and to the left from $(\hat{\pi}, \hat{m})$ and is determined by the borderline case where the first order condition

for the minimisation problem in (13) holds for $q = q_{\max}$. Similarly, the rightmost ray which separates the last two regions goes up and to the right and is determined by the borderline case where the first order condition holds for $q = q_{\min}$. Interior solutions are obtained in the regions (denoted by $\mathcal{A}_{\text{int},\ell}$ and $\mathcal{A}_{\text{int},r}$) which lie below these rays, and are given by[17]

$$O(\pi, v) = q^m(\pi) + \frac{v - m(\pi)}{m(\pi) - \hat{m}} \left[q^m(\pi) - \hat{q}\right] . \tag{15}$$

Note that this is the myopic quantity plus an adjustment away from $\hat{q}$, i.e. in the direction of more informative price signals. Evaluating the minimand in (13) at these quantities, we find

$$G(\pi, v) = \beta(\pi) \frac{v - m(\pi)}{v - \hat{m}} \tag{16}$$

in the two regions associated with interior solutions.

Corner solutions are obtained in the area between the leftmost and rightmost rays. This area splits into two regions along a third, central ray (denoted by $\mathcal{R}_c$) which is determined by the borderline case when $q_{\max}$ and $q_{\min}$ are *both* optimal and so give the same value in (13). We have $O(\pi, v) = q_{\max}$ between the left and the central ray, $O(\pi, v) = \{q_{\max}, q_{\min}\}$ along the central ray, and $O(\pi, v) = q_{\min}$ between the central and the right ray. The corresponding expressions for G are

$$G(\pi, v) = \frac{v - R(\pi, q_{\max})}{[q_{\max} - \hat{q}]^2} \quad \text{and} \quad G(\pi, v) = \frac{v - R(\pi, q_{\min})}{[\hat{q} - q_{\min}]^2} . \tag{17}$$

Given the representations (16) – (17) on the respective regions, it is now straightforward to verify that G is continuous on $\mathcal{A}$.[18]

3.5 The Adjusted Value Function

The above results show that optimal quantities depend only on the graph of the function

$$v^*(\pi) = u^*(\pi) - \lambda(\pi) \frac{(u^*)'(\pi)}{r} \tag{18}$$

which we call the *adjusted value function* (adjusted for the contribution of state switching). Since knowing the adjusted value function will be enough to determine optimal policies, our next step is to transform the ODE (14) for u^* into an ODE for

[17] When $O(\pi, v)$ is a singleton, we write $O(\pi, v) = q$ rather than $O(\pi, v) = \{q\}$.

[18] Note, however, that it cannot be extended continuously into the point $(\hat{\pi}, \hat{m})$ that we excluded from the set $\mathcal{A}$. For a sequence of points (π_n, v_n) in $\mathcal{A}$ converging to $(\hat{\pi}, \hat{m})$ along the graph of the myopic pay-off function m, for example, $\lim_{n\to\infty} G(\pi_n, v_n) = 0$; for a sequence converging to $(\hat{\pi}, \hat{m})$ along the central ray, on the other hand, this limit is $\beta(\hat{\pi})$.

v^*. Note that the boundary conditions become very simple: v^* and m coincide at the beliefs 0 and 1. Note also that once we know v^*, we can recover u^* by integrating (18), that is, by solving a linear ODE.[19]

We formally differentiate the equation $v(\pi) = u(\pi) - \lambda(\pi)\, u'(\pi)/r$ twice, each time using the relationship $\tau(\pi)u''(\pi)/2r = G(\pi, v(\pi))$ to replace u'' with an expression that involves only π and v. This yields the following second-order ODE for the adjusted value function:

$$\tau(\pi)\frac{v''(\pi)}{2} = r\, G(\pi, v(\pi)) + \Lambda \left\{ f(\pi)\, G(\pi, v(\pi)) + (\pi - \tilde{\pi})\frac{d}{d\pi} G(\pi, v(\pi)) \right\} \tag{19}$$

with

$$f(\pi) = 2 - (\pi - \tilde{\pi})\frac{\tau'(\pi)}{\tau(\pi)} = 2\left(\frac{\tilde{\pi}(1-\pi)}{\pi} + \frac{(1-\tilde{\pi})\pi}{1-\pi} \right).$$

As for differentiability of G, it is easy to check that G is continuously differentiable in the interior of $\mathcal{A}$ with the exception of the central ray separating the regions where $q_{\max}$ or $q_{\min}$ is optimal. We will therefore consider the ODE (19) separately to the left and to the right of that ray.

Summarising the developments so far, we can say that the adjusted value function solves (19) on $]0,1[$ with the possible exception of the confounding belief $\hat{\pi}$ or any belief where the graph of v^* crosses the central ray. Conversely, if we have a solution (in the sense of the previous sentence) v of (19) with the above boundary conditions and such that $q^*(\pi) = O(\pi, v(\pi))$ defines an admissible policy function, then $v = v^*$ and the policy q^* is optimal. It is mainly this version of a verification theorem that we will use below.

When the optimisation problem in the Bellman equation has an interior solution, $G(\pi, v)$ is given by (16), so the ODE (19) becomes

$$\begin{aligned} \tau(\pi)\frac{v''(\pi)}{2} \;=\; & \beta(\pi) \left\{ \left(r + \Lambda \left[f(\pi) + (\pi - \tilde{\pi})\frac{\Delta\beta}{\beta(\pi)} \right] \right) \frac{v(\pi) - m(\pi)}{v(\pi) - \hat{m}} \right. \\ & \left. + \Lambda\, (\pi - \tilde{\pi}) \left[\frac{v(\pi) - m(\pi)}{v(\pi) - \hat{m}} \right]' \right\} \end{aligned} \tag{20}$$

in this case. Many of the results obtained in the following sections are based on a detailed investigation of this particular differential equation.

[19] In fact, the method of variation of constants shows that for $\Lambda > 0$ and a continuous function v,

$$u(\pi) = (r/\Lambda)\, |\pi - \tilde{\pi}|^{-r/\Lambda} \operatorname{sign}(\pi - \tilde{\pi}) \int_{\tilde{\pi}}^{\pi} |\xi - \tilde{\pi}|^{r/\Lambda - 1}\, v(\xi)\, d\xi$$

is the unique bounded solution of the ODE $u(\pi) - \lambda(\pi)\, u'(\pi)/r = v(\pi)$.

While we have derived the above statements for a discount rate $r > 0$ only, they continue to be valid in the limiting case of no discounting ($r = 0$) once we use a definition of the adjusted value function that corresponds to the so-called *catching-up criterion*; we refer the reader to Appendix E for details.[20] The undiscounted case provides a useful benchmark; in fact, economic intuition suggests that the optimal experimentation strategy of an agent with discount rate $r > 0$ will be 'in between' the two extremes given by myopic behaviour (corresponding to $r = \infty$), on the one hand, and the behaviour of an infinitely patient agent ($r = 0$), on the other hand.[21]

4 Experimentation Regimes

Our assumption that the confounding quantity $\hat{q}$ lies in the interior of Q^m brings with it two complications. A first complication arises from the fact that it might be optimal to choose $\hat{q}$ at $\hat{\pi}$. As we have already seen, choosing $\hat{q}$ leads to a completely uninformative price signal, makes the diffusion coefficient in the updating equation (2) vanish and thereby causes a singularity in the Bellman equation and the related ODEs. Moreover, there is a 'break' in the ODE for the adjusted value function along the central ray.

A second complication arises from the fact that the direction of increasing informativeness of the price signal is ambiguous. Assume for example that the current belief is slightly higher than $\hat{\pi}$, so the myopically optimal quantity is slightly below $\hat{q}$. The true optimum will usually involve some deviation from the myopic quantity, motivated by the desire to render observed prices more informative, and naïve intuition suggests that the monopolist might wish to move further away from $\hat{q}$ by reducing quantity. However, it could also make sense to *increase* quantity beyond $\hat{q}$ and thus achieve a wider spread between the two possible price distributions there. For beliefs close to the boundaries of the unit interval, on the other hand, we do expect the naïve intuition to be borne out. Thus, we expect optimal experimentation to involve quantity expansion for beliefs π close to 0, and quantity reduction for beliefs π close to 1. The optimal policy as a function of beliefs will then have to move downward past $\hat{q}$ as π increases, and, in doing so, will either select $\hat{q}$ at $\hat{\pi}$, or avoid $\hat{q}$ altogether by jumping past it.

Confirming this intuition, we shall find two different regimes of optimal exper-

[20] See Dutta (1991) for a discussion of undiscounted decision criteria in a discrete time framework.

[21] Moreover, it is well known that the undiscounted case tends to be mathematically more tractable than the discounted case. See for instance Bolton and Harris (1993) and Harris (1988).

imentation. In the *moderate* experimentation regime, the optimal policy selects quantities only in Q^m, the range of the myopic policy, and it selects $\hat{q}$ at $\hat{\pi}$. In the *extreme* experimentation regime, each of the quantities $q_{\max}$ and $q_{\min}$ is chosen on a set of beliefs of positive measure; in particular, $q_{\max}$ or $q_{\min}$ will be chosen at $\hat{\pi}$, and the optimal policy will exhibit a jump past $\hat{q}$ from one extreme quantity to the other. These regimes are further distinguished by the sample path behaviour of posterior beliefs and optimal quantities. While extreme experimentation implies that any posterior belief can be reached with positive probability, moderate experimentation restricts posterior beliefs to lie on one side of $\hat{\pi}$ in the long run, so the monopolist ends up producing quantities from only part of Q^m.

After characterising the adjusted value function and the optimal policy in the two regimes, we will show that extreme experimentation arises for low values of r, Λ and σ, and moderate experimentation for high values. Near the boundary between the corresponding parameter regions, a small change in any of these parameters can trigger a change in the experimentation regime, hence a large discontinuous change in the monopolist's strategy and the resulting sample path behaviour of beliefs and quantities produced. These results are particularly clear in the limiting cases where at least one of the parameters r and Λ is zero. In the undiscounted case with state switching ($r = 0$ and $\Lambda > 0$) we will establish the existence of a critical switching intensity that separates moderate from extreme experimentation. Similarly, we will find a critical discount rate in the discounted case without state switching ($r > 0$ and $\Lambda = 0$). Finally, the simple benchmark where the monopolist is infinitely patient ($r = 0$) and the environment does not change ($\Lambda = 0$) allows a closed-form solution for the optimal policy.

Throughout this section, we fix demand curve parameters α_0, α_1, β_0 and β_1 such that the confounding quantity $\hat{q}$ lies in the interior of Q^m; an invariant belief $\tilde{\pi} \in]0, 1[- \{\hat{\pi}\}$ is also held fixed.

4.1 Moderate versus Extreme Experimentation

The discussion in Section 3.2 suggests that the monopolist's behaviour will depend crucially on whether the inequality $v^*(\hat{\pi}) \geq \hat{m}$ for the adjusted value function at the confounding belief is strict or holds with equality. The following theorem characterises the adjusted value function in either case. It is the first step towards a description of the corresponding optimal behaviour.

Let $(D_{\tilde{\pi}} v)(\hat{\pi})$ denote the one-sided derivative of a function v at $\hat{\pi}$ in the direction of $\tilde{\pi}$, and $(D_o v)(\hat{\pi})$ the one-sided derivative in the opposite direction. Recall the

structure of the ODE (19) for the adjusted value function, and in particular its special case (20) associated with interior solutions of the optimisation problem in the Bellman equation.

Theorem 4.1 *If $v^*(\hat{\pi}) > \hat{m}$, then the adjusted value function is the unique differentiable function which solves the ODE* (19) *on $\{\pi \in]0,1[: (\pi, v^*(\pi)) \notin \mathcal{R}_c\}$ subject to $v^*(\pi) = m(\pi)$ at $\pi = 0,\ 1$ and $v^* > m$ on $]0,1[$.*

If $v^*(\hat{\pi}) = \hat{m}$, on the other hand, then the adjusted value function is the unique solution of the ODE (20) *on $]0,1[\ -\{\hat{\pi}\}$ subject to $v^*(\pi) = m(\pi)$ at $\pi = 0,\ \hat{\pi},\ 1$, $v^* > m$ on $]0,1[\ -\{\hat{\pi}\}$, and $(D_{\tilde{\pi}} v^*)(\hat{\pi}) = 0$; moreover, it is strictly convex with $(v^*)'' > 0$ on $]0,1[\ -\{\hat{\pi}\}$.*

> Note that the statement for $v^*(\hat{\pi}) = \hat{m}$ does not say anything about the one-sided derivative $(D_o v^*)(\hat{\pi})$. This allows for the possibility that $(D_o v^*)(\hat{\pi}) \neq 0$ and hence for the adjusted value function to have a kink at $\hat{\pi}$.

PROOF: For $v^*(\hat{\pi}) = \hat{m}$, the statements on convexity and the one-sided derivative $(D_{\tilde{\pi}} v^*)(\hat{\pi})$ are shown in Appendix F; see Proposition F.2. Given $v^*(\hat{\pi}) = \hat{m}$ and the boundary conditions, convexity entails $(\pi, v^*(\pi)) \in \mathcal{A}_{\text{int},\ell} \cup \mathcal{A}_{\text{int},r}$ for all $\pi \in]0,1[\ -\{\hat{\pi}\}$. Sections 3.3 and 3.5 therefore imply that v^* solves (20) on $]0,1[\ -\{\hat{\pi}\}$ subject to the stated conditions. If v is another solution of (20) on $]0,1[\ -\{\hat{\pi}\}$ with $v(\pi) = m(\pi)$ at $\pi = 0,\ \hat{\pi},\ 1$ and $(D_{\tilde{\pi}} v)(\hat{\pi}) = 0$, then the construction of optimal strategies in the proof of Proposition 4.2 below and our verification theorem imply that $v = v^*$. Finally, if there were a $\pi \neq 0,\ \hat{\pi},\ 1$ such that $v^*(\pi) = m(\pi)$, then $v^* - m$ would have a local minimum there, hence $(v^*)''(\pi) \geq m''(\pi) > 0$. But $v^*(\pi) = m(\pi)$ would imply $(u^*)''(\pi) = 0$ by ODE (14) (or its undiscounted variant), hence $(u^*)''$ would have a local minimum at π. This would require $(u^*)'''(\pi) = 0$ and, since $(v^*)''(\pi)$ is a linear combination of $(u^*)''(\pi)$ and $(u^*)'''(\pi)$, we would have $(v^*)''(\pi) = 0$ – a contradiction.

For $v^*(\hat{\pi}) > \hat{m}$, Sections 3.3 and 3.5 imply that v^* is once continuously differentiable and solves the ODE (19) on $\{\pi \in]0,1[: (\pi, v^*(\pi)) \notin \mathcal{R}_c\}$. Given another solution v with this property and the same boundary conditions, the arguments in the proof of Proposition 4.1 below together with the verification theorem imply again that $v = v^*$. Finally, the same argument as above shows $v^* > m$ for all beliefs $\pi \neq 0,\ 1$ with $(\pi, v^*(\pi)) \notin \mathcal{R}_c$; this implies $v^* > m$ on the whole of $]0,1[$. ∎

The following two propositions describe the optimal behaviour for $v^*(\hat{\pi}) > \hat{m}$ and $v^*(\hat{\pi}) = \hat{m}$, respectively.

Proposition 4.1 (Extreme Experimentation) *If $v^*(\hat{\pi}) > \hat{m}$, then the optimal policy function prescribes each of the extreme quantities $q_{\max}$ and $q_{\min}$ on a set of*

beliefs of positive measure, one of which contains $\hat{\pi}$, and it is continuous except for a jump from one extreme quantity to the other at any belief π such that $(\pi, v^*(\pi)) \in \mathcal{R}_c$. The corresponding process of posterior beliefs is regular on $]0,1[$, that is, starting from any point in this interval, any other point in it may be reached with positive probability.

PROOF: The policy function q^* obtained by extending $O(\pi, v^*(\pi))$ continuously into $\pi = 0$ and 1 and selecting either $q_{\max}$ or $q_{\min}$ at any π where $(\pi, v^*(\pi)) \in \mathcal{R}_c$ is piecewise continuous with $q^*(\pi) - \hat{q}$ bounded away from zero, hence admissible. Optimality now follows from the verification theorem in Section 3.5. Clearly, $q^*(\pi) = q_{\max}$ on a set of positive measure, and the same is true for $q_{\min}$. It is also clear that one of these sets contains $\hat{\pi}$. The fact that $q^*(\pi) - \hat{q}$ is bounded away from zero also implies regularity of the process of posterior beliefs. ■

Whenever $v^*(\hat{\pi}) > \hat{m}$, we thus find extreme experimentation in the sense that the quantities $q_{\max}$ and $q_{\min}$ are optimal for non-negligible sets of beliefs. Moreover, optimal quantities are always some distance away from the confounding quantity, so the information content of the price signal observed by the monopolist is bounded away from zero. The resulting process of posterior beliefs can therefore reach any point in the open unit interval with positive probability.

An example of extreme experimentation is shown in Figure 3.4 which has been calculated for $r = 0.1$ and $\Lambda = 0.05$.[22] The bold line in the upper panel is the graph of the adjusted value function v^*, the thin line that of the myopic pay-off function m, and the thick grey line that of the value function u^*. (The upper panel also shows the three rays introduced in Section 3.4.) In the lower panel, the bold line is the optimal policy function q^*, while the thin line is the myopic policy q^m. The adjusted value is strictly higher than the myopic pay-off at all non-degenerate beliefs, and as v^* crosses each ray in turn, q^* first reaches $q_{\max}$, then jumps to $q_{\min}$, and finally moves away from $q_{\min}$; the jump occurs after $\hat{\pi}$ because the central ray goes up and to the right for the demand curve parameters and the range of feasible quantities used in this example. Note that for beliefs between $\hat{\pi}$ and the jump, the monopolist generates a more informative price signal by choosing q^* on the side of $\hat{q}$ *opposite* to where q^m lies.

[22] In this and all the subsequent figures at the end of the chapter, the demand curve parameters are $\alpha_0 = 40$, $\beta_0 = 2/3$, $\alpha_1 = 60$ and $\beta_1 = 3/2$, implying $q^m(0) = 30$, $q^m(1) = 20$, $\hat{q} = 24$ and $\hat{\pi} = 0.4$. The range of feasible quantities is defined by $q_{\min} = 40/3$ and $q_{\max} = 40$, the noise parameter is $\sigma = 5$, and the invariant belief is $\tilde{\pi} = 0.5$. Only r and Λ vary across figures. The adjusted value function is calculated as a numerical solution to the ODE (19) subject to the boundary conditions $v^*(0) = m(0)$ and $v^*(1) = m(1)$, and optimal quantities are then determined through the optimal policy correspondence. Details of the numerical procedure are reported in Appendix H.

The corresponding sample path behaviour is illustrated in Figure 3.5. The upper panel shows the evolution of the agent's belief starting from the prior $\pi_0 = 0.25$, the lower panel the associated quantities. The bold dashed line in either panel represents the true state, the initial state being $k_0 = 0$. By the time of the first state change, the agent's belief has predominantly been between 0 and 0.2. After the state change, the belief moves relatively quickly in the direction of the new state and eventually reaches $\hat{\pi}$. This starts a phase of intense experimentation with frequent jumps between $q_{\max}$ and $q_{\min}$. At the end of this phase, the belief leaves the neighbourhood of $\hat{\pi}$ to move closer to the true state. This pattern is repeated each time the state switches, and the true state is tracked quite well. Observe that the relatively stable environment induces high variability in the agent's actions.

Next, we turn to the case where $v^*(\hat{\pi}) = \hat{m}$. This is the more complicated case since it involves the 'singularity' at $(\hat{\pi}, \hat{m})$ of the ODE for v^*. We formulate the next result for $\Lambda > 0$; we shall obtain the analogous result for $\Lambda = 0$ later, in Section 4.3.

Proposition 4.2 (Moderate Experimentation) *Let $\Lambda > 0$ and $v^*(\hat{\pi}) = \hat{m}$. Then the optimal policy assumes values in Q^m only and selects $\hat{q}$ at $\hat{\pi}$. With probability one, the resulting process of posterior beliefs is, in the long run, confined to the subinterval $]0, \hat{\pi}[$ or $]\hat{\pi}, 1[$ which contains $\tilde{\pi}$, and the monopolist ends up choosing quantities in either $]\hat{q}, q^m(0)[$ or $]q^m(1), \hat{q}[$ only.*

> More precisely, the proof will show that the optimal policy function is differentiable if v^* is differentiable at $\hat{\pi}$, while it has a single jump at $\hat{\pi}$ if v^* has a kink.

PROOF: For $\pi \in]0,1[- \{\hat{\pi}\}$, we have $(\pi, v^*(\pi)) \in \mathcal{A}_{\text{int},\ell} \cup \mathcal{A}_{\text{int},r}$ by convexity of v^* and hence

$$O(\pi, v^*(\pi)) = q^m(\pi) + \frac{v^*(\pi) - m(\pi)}{m(\pi) - \hat{m}} \, [q^m(\pi) - \hat{q}] = \hat{q} + \frac{v^*(\pi) - \hat{m}}{m(\pi) - \hat{m}} \, [q^m(\pi) - \hat{q}]$$

from (15). Strict convexity of v^* also implies that $v^* < \overline{m}_\ell$ on $]0, \hat{\pi}[$ and $v^* < \overline{m}_r$ on $]\hat{\pi}, 1[$ with $\overline{m}_\ell$ and $\overline{m}_r$ being the functions whose graphs are the straight lines joining the point $(\hat{\pi}, \hat{m})$ with the points $(0, m(0))$ and $(1, m(1))$, respectively. It is straightforward to verify that $O(\pi, \overline{m}_\ell(\pi)) = q^m(0)$ on $]0, \hat{\pi}[$ and $O(\pi, \overline{m}_r(\pi)) = q^m(1)$ on $]\hat{\pi}, 1[$. Since $v^* > m$ on $]0,1[- \{\hat{\pi}\}$, we conclude that $q^m(\pi) < O(\pi, v^*(\pi)) < q^m(0)$ on $]0, \hat{\pi}[$ and $q^m(\pi) > O(\pi, v^*(\pi)) > q^m(1)$ on $]\hat{\pi}, 1[$. In particular, $O(\pi, v^*(\pi))$ assumes values in Q^m only.

Straightforward algebra shows that

$$O(\pi, v^*(\pi)) = \hat{q} - \frac{2}{\Delta\alpha} \, \frac{v^*(\pi) - \hat{m}}{\pi - \hat{\pi}} \, ,$$

so $O(\pi, v^*(\pi)) \to \hat{q}$ as $\pi \to \hat{\pi}\pm$ iff $(D_\pm v^*)(\hat{\pi}) = 0$.

If $(v^*)'(\hat{\pi}) = 0$, $O(\pi, v^*(\pi))$ can therefore be extended to a continuous policy function $q^* : [0,1] \to Q^m$ with $q^*(\hat{\pi}) = \hat{q}$. In fact, the policy is differentiable with bounded derivative, hence Lipschitz continuous. This is obvious for beliefs different from $\hat{\pi}$; differentiability at $\hat{\pi}$ follows from the representation

$$\frac{q^*(\pi) - \hat{q}}{\pi - \hat{\pi}} = \frac{v^*(\pi) - \hat{m}}{m(\pi) - \hat{m}} \; \frac{q^m(\pi) - \hat{q}}{\pi - \hat{\pi}}$$

and the fact that the ratio $[v^*(\pi) - \hat{m}]/[m(\pi) - \hat{m}]$ tends to a finite limit as $\pi \to \hat{\pi}$ (see Proposition F.2). The policy q^* is admissible by Proposition A.1, hence optimal by the verification theorem from Section 3.5. Turning to the belief process resulting from this policy, let us assume $\tilde{\pi} > \hat{\pi}$ for concreteness. Starting from a prior belief π_0 in the subinterval $[\hat{\pi}, 1]$, all posterior beliefs π_t will remain in the open subinterval $]\hat{\pi}, 1[$ because the inequality $\lambda(\hat{\pi}) > 0$ makes the belief $\hat{\pi}$ an entrance boundary. (Since $\lambda(1) < 0$ for $\tilde{\pi} \neq 1$, the right boundary of the unit interval is always an entrance boundary.) If $\pi_0 < \hat{\pi}$, on the other hand, the process of beliefs will, with probability one, reach $\hat{\pi}$ in finite time and then move into the subinterval $]\hat{\pi}, 1[$.

Next, suppose that v^* has a kink at $\hat{\pi}$. To be concrete, we assume again that $\tilde{\pi} > \hat{\pi}$, so $(D_- v^*)(\hat{\pi}) < 0$ and $(D_+ v^*)(\hat{\pi}) = 0$. Let q^* be the policy function obtained by extending $O(\pi, v^*(\pi))$ continuously into $\pi = 0$ and 1 and setting $q^*(\hat{\pi}) = \hat{q}$; again, this function takes values in Q^m only. By Proposition F.2, the ratio $[v^*(\pi) - \hat{m}]/[m(\pi) - \hat{m}]$ tends to a finite limit as $\pi \to \hat{\pi}+$. As above, this implies that the restriction of the policy function q^* to the interval $[\hat{\pi}, 1]$ is Lipschitz continuous, hence admissible. Moreover, if the prior belief π_0 lies in this subinterval and the monopolist uses the policy q^*, then all posterior beliefs π_t will remain in $]\hat{\pi}, 1[$ by the same argument as above. By the verification theorem, the policy q^* is thus optimal for all prior beliefs $\pi_0 \geq \hat{\pi}$.

From the above expression for $O(\pi, v^*(\pi))$, we see that q^* approaches the limit

$$q^*(\hat{\pi}-) = \hat{q} - \frac{2}{\Delta\alpha} (D_- v^*)(\hat{\pi}) \in]\hat{q}, q^m(0)]$$

from the left of $\hat{\pi}$. On the subinterval $[0, \hat{\pi}[$, the function q^* is (locally) Lipschitz continuous, so the existence result underlying Proposition A.1 implies that, starting from any prior belief $\pi_0 < \hat{\pi}$, the policy q^* generates a unique stochastic process of beliefs up to the first time $\hat{\pi}$ is reached; with probability one, this happens in finite time. From then on, the process of beliefs is uniquely determined by the restriction of q^* to $[\hat{\pi}, 1]$. This establishes admissibility of the policy q^* on the whole of the unit interval,[23] and optimality follows

[23]The technical details required to make this argument fully rigorous are beyond the scope of this chapter. One can avoid these complications altogether without affecting the main results by constructing ϵ-optimal policies. Given an arbitrary $\epsilon > 0$, define $\delta = \epsilon/(1 + \max_{q \in Q^m} [q - \hat{q}]^2)$. Since $(D_- v^*)(\hat{\pi}) < 0$, $\tau(\pi)(u^*)''(\pi)/(2r)$ converges to $\beta(\hat{\pi})$ as $\pi \to \hat{\pi}-$. (We only deal with the

from the verification theorem. ∎

The proposition shows that optimal experimentation is moderate whenever $v^*(\hat{\pi}) = \hat{m}$, meaning that the monopolist restricts himself to quantities in Q^m. The optimal policy function approaches the confounding quantity from at least one side of $\hat{\pi}$, the side where $\tilde{\pi}$ lies. In a changing environment, this implies that starting from any prior belief lying on the same side of $\hat{\pi}$ as $\tilde{\pi}$, the process of posterior beliefs will stay on this side forever; starting from a prior belief on the other side of $\hat{\pi}$, the process of posterior beliefs will cross $\hat{\pi}$ almost surely in finite time and then be confined to the side where $\tilde{\pi}$ lies. Eventually, the monopolist's beliefs will thus be 'trapped' on one side of $\hat{\pi}$, although he knows that the state of demand will continue to switch from time to time. This result is the analogue, in a changing environment, of the possibility of cessation of learning in an unchanging environment as identified by Rothschild (1974), McLennan (1984), Easley and Kiefer (1988), Aghion, Bolton, Harris and Jullien (1991) and others; cf. our discussion of the case $\Lambda = 0$ below.

Figure 3.6 shows an example of moderate experimentation, calculated for $r = 0.1$ and $\Lambda = 0.2$. Again, the bold line in the upper panel is the adjusted value function v^*, the thin line the myopic pay-off function m, and the thick grey line the value function u^*. In the lower panel, the bold line is the optimal policy function q^*, and the thin line the myopic policy q^m. The adjusted value function v^* touches the myopic pay-off at its lowest point. Consequently, the optimal policy q^* never selects quantities outside the range of the myopic policy, spanned by $q^m(0) = 30$ and $q^m(1) = 20$. (For this reason, the vertical axis is scaled differently from that in Figure 3.4.) Note that q^* always lies on the same side of $\hat{q}$ as q^m, but further away. Furthermore, it appears that v^* is differentiable at $\hat{\pi} = 0.4$, and q^* moves smoothly through $\hat{q} = 24$.

Figure 3.7 illustrates the corresponding sample path behaviour. The upper panel shows how the belief process is trapped after its transit through $\hat{\pi}$; in particular, the true state (again represented by a bold dashed line) is tracked poorly. As a

---

case $r > 0$ here; a similar argument can be given for the undiscounted case.) Next, the continuity of v^* implies that $R(\pi, q) - v^*(\pi)$ converges to $-\beta(\hat{\pi})[q - \hat{q}]^2$ as $\pi \to \hat{\pi}$, and this convergence is uniform in $q \in Q^m$. So we can find $\rho > 0$ such that $\tau(\pi)(u^*)''(\pi)/(2r) \geq \beta(\hat{\pi}) - \delta$ and $R(\pi, q) - v^*(\pi) \geq -\beta(\hat{\pi})[q - \hat{q}]^2 - \delta$ for all $\pi \in [\hat{\pi} - \rho, \hat{\pi}]$ and all $q \in Q^m$. The Lipschitz continuous (hence admissible) policy function $q_\epsilon : [0, 1] \to Q^m$ which coincides with q^* on $[0, \hat{\pi} - \rho] \cup [\hat{\pi}, 1]$ and whose graph joins the points $(\hat{\pi} - \rho, q^*(\hat{\pi} - \rho))$ and $(\hat{\pi}, \hat{q})$ by a straight line satisfies

$$\tau(\pi)\frac{(u^*)''(\pi)}{2r}\,[q_\epsilon(\pi) - \hat{q}]^2 + R(\pi, q_\epsilon(\pi)) - v^*(\pi) \geq -\epsilon$$

on $[\hat{\pi} - \rho, \hat{\pi}]$, hence is ϵ-optimal by Proposition C.2. The resulting long-run behaviour of beliefs and actions is as in the proposition.

consequence, the resulting path of optimal quantities in the lower panel remains in the range from $q^m(1) = 20$ to $\hat{q} = 24$ after its first passage through the confounding quantity. (Again note that the vertical axis is scaled differently from that in Figure 3.5.) Observe that with a more unstable environment, the agent's actions become less variable.

The moderate experimentation scenario where v^* has a kink at $\hat{\pi}$ is particularly interesting.[24] If $\tilde{\pi} > \hat{\pi}$, say, and $(D_- v^*)(\hat{\pi}) > 0$, the optimal policy approaches a limit different from $\hat{q}$ as $\pi \to \pi-$; see the proof of Proposition 4.2. This is due to the fact that to the left of $\hat{\pi}$, $(v^*)'$ is bounded away from zero, so $(u^*)''$ is relatively high, implying a high value of information. Intuitively, we can interpret this as follows. With a posterior belief slightly to the left of $\hat{\pi}$, the agent anticipates that once his belief crosses $\hat{\pi}$, he will not find it profitable to experiment in a way that would allow his belief to cross $\hat{\pi}$ from the right to the left again. Therefore, he experiments relatively strongly so as to give his belief a chance to avoid the trap for now and move away from $\hat{\pi}$ to the left, should the true state currently be $k = 0$.

In Figure 3.8, there is an example of moderate experimentation when $r = 0.1$ and Λ has the intermediate value 0.15. While the kink in the adjusted value function is difficult to detect visually, the shape of the graph of the optimal policy[25] immediately to the left of $\hat{\pi}$ is clearly somewhat different from that in Figure 3.6. The corresponding sample path behaviour illustrated in Figure 3.9 is quite similar to that in Figure 3.7 once the belief has transited $\hat{\pi}$, but, prior to that, there is a noticeable difference.

Whether experimentation is moderate or extreme depends on the underlying parameter combination (r, Λ, σ). In fact, the parameter space $\mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}_{++}$ splits into two sets, one where $v^*(\hat{\pi}) = \hat{m}$ and experimentation is moderate, and another where $v^*(\hat{\pi}) > \hat{m}$ and experimentation is extreme. The above numerical examples suggest (and we shall prove below) that neither set is empty. Thus, the optimal strategy and the resulting long-run behaviour of beliefs and actions depend *qualitatively* on the parameters r, Λ and σ. Moreover, a small change in one of these parameters can trigger a large discontinuous change in the monopolist's strategy and the resulting stream of quantities that he produces. We will see this particularly clearly in the special cases which we study below.

[24] Note that this is the only case where the value function u^* fails to be twice continuously differentiable on the whole of $]0, 1[$.

[25] Because of the inherent smoothing nature of the numerical procedure, the data has been slightly re-adjusted and the graph should be thought of as illustrating an ϵ-optimal policy.

4.2 Sufficient Conditions and Critical Parameter Values

We would naturally expect moderate experimentation for high values of r, Λ and σ, and extreme experimentation for low values. As the discount rate increases, for example, the future becomes less important to the agent, the value of information falls, and with it the agent's willingness to sacrifice current revenue for potential future gains from experimentation. A higher level of noise, on the other hand, renders the price signal less informative, which reduces the expected gain from any given deviation from the myopic optimum, and thus the incentive to experiment. Last, a higher frequency of state switches increases the risk of information becoming obsolete, so the trade-off between current revenues and potential gains from experimentation again shifts in favour of the former. For high values of these parameters, therefore, the monopolist rationally assesses the loss in current revenue from experimenting strongly near $\hat{\pi}$ to be higher than the loss in future revenues resulting from sometimes being trapped on the 'wrong' side of $\hat{\pi}$. Before formulating a result to this effect, we first note that the choice of the interval Q of feasible quantities is irrelevant here.

Corollary 4.1 *Given a parameter combination (r, Λ, σ), experimentation is moderate (extreme) for* some $Q \supseteq Q^m$ *if and only if it is moderate (extreme) for* all $Q \supseteq Q^m$.

PROOF: This follows directly from Theorem 4.1 since the ODE (20), which characterises v^* when experimentation is moderate, does not involve the quantities $q_{\max}$ and $q_{\min}$. ∎

We are therefore free to choose the interval Q in a convenient way when we look for sufficient conditions for either type of experimentation. This is exploited in the proof of the following result.

Proposition 4.3 *There are positive constants $\rho_m \geq \rho_e$ and $\ell_m \geq \ell_e$ such that optimal experimentation is moderate with a differentiable policy function for all $r \geq 0$, $\Lambda \geq 0$ and $\sigma > 0$ satisfying*

$$\frac{r}{\rho_m} + \frac{\Lambda}{\ell_m} \geq \frac{1}{\sigma^2},$$

and extreme for all $r \geq 0$, $\Lambda \geq 0$ and $\sigma > 0$ satisfying

$$\frac{r}{\rho_e} + \frac{\Lambda}{\ell_e} \leq \frac{1}{\sigma^2}.$$

PROOF: By Proposition G.3, there are positive constants c_1 and c_2 such that for all $r \geq 0$, $\Lambda \geq 0$ and $\sigma > 0$ satisfying $r + c_1\Lambda \geq c_2/\sigma^2$, there exists a continuous function $v : [0,1] \to \mathbb{R}$ which solves (20) on $]0,1[- \{\hat{\pi}\}$ with $v(\pi) = m(\pi)$ at $\pi = 0,\ \hat{\pi},\ 1$ and $v > m$ everywhere else; in particular, v is differentiable at $\hat{\pi}$ with $v'(\hat{\pi}) = 0$. The uniqueness part of Theorem 4.1 for $v^*(\hat{\pi}) = \hat{m}$ implies that $v = v^*$. Experimentation is therefore moderate with a continuous optimal policy for all these parameter combinations. We can thus set $\rho_m = c_2$ and $\ell_m = c_2/c_1$.

Turning to extreme experimentation, we assume without loss of generality that Q is centred on $\hat{q}$. (This makes the central ray $\mathcal{R}_c$ vertical and simplifies the construction of solutions to the ODE (19) via the techniques of Appendix G.) For this case, Proposition G.4 shows that there are positive constants c_3, c_4 and c_5 such that for all $r \geq 0$, $\Lambda \geq 0$ and $\sigma > 0$ satisfying $c_3 r + c_4\Lambda \leq c_5/\sigma^2$, there exists a continuous function $v : [0,1] \to \mathbb{R}$ which is once continuously differentiable and which solves (19) on $]0,1[- \{\hat{\pi}\}$ with $v(\pi) = m(\pi)$ at $\pi = 0,\ 1$ and $v > m$ everywhere else. By the uniqueness part of Theorem 4.1 for $v^*(\hat{\pi}) > \hat{m}$, we have $v = v^*$, so experimentation is extreme for these parameter combinations, and we can take $\rho_e = c_5/c_3$ and $\ell_e = c_5/c_4$. ∎

Note that the constructive approach used in the proof (based on Propositions G.3 and G.4) yields explicit formulae for the constants ρ_m, ℓ_m, ρ_e and ℓ_e. The proof also shows that under the stated condition for moderate experimentation, the adjusted value function is always differentiable at $\hat{\pi}$. In particular, a kink in v^* can only occur in an 'intermediate' range of parameter combinations.

Using the fact that v^* is strictly convex whenever $v^*(\hat{\pi}) = \hat{m}$, we can derive a more precise characterisation of the boundary between the parameter regions associated with moderate and extreme experimentation. Before formulating this result, we note from the ODE (19) that v^* depends on r, Λ and σ only through the two products $\rho = r\sigma^2$ and $\ell = \Lambda\sigma^2$. This reduces the parameter space effectively to the non-negative orthant $\mathbb{R}^2_+$, which splits into a region of moderate experimentation and a region of extreme experimentation. As the following result shows, the boundary between these two regions cuts each ray through the origin in a single point; see Figure 3.3. Thus, we can 'trace' this boundary by varying the slope of the ray.

Proposition 4.4 *Let R be a ray in $\mathbb{R}^2_+$ emanating from the origin $(0,0)$. Then, there is a unique point $(\rho^\dagger, \ell^\dagger) \in R - \{(0,0)\}$ such that experimentation is extreme for all $(\rho,\ell) \in R$ with $(\rho,\ell) < (\rho^\dagger,\ell^\dagger)$, and moderate for all $(\rho,\ell) \in R$ with $(\rho,\ell) \geq (\rho^\dagger,\ell^\dagger)$.*

PROOF: Fixing a point $(\rho_0,\ell_0) \in R - \{(0,0)\}$, we can parameterise the ray R by the mapping $\mu \mapsto (\mu\rho_0, \mu\ell_0)$ where $\mu \geq 0$. We write $v^*[\mu]$ for the adjusted value function associated with the parameters $(\mu\rho_0, \mu\ell_0)$, and ODE$[\mu]$ for the corresponding differential

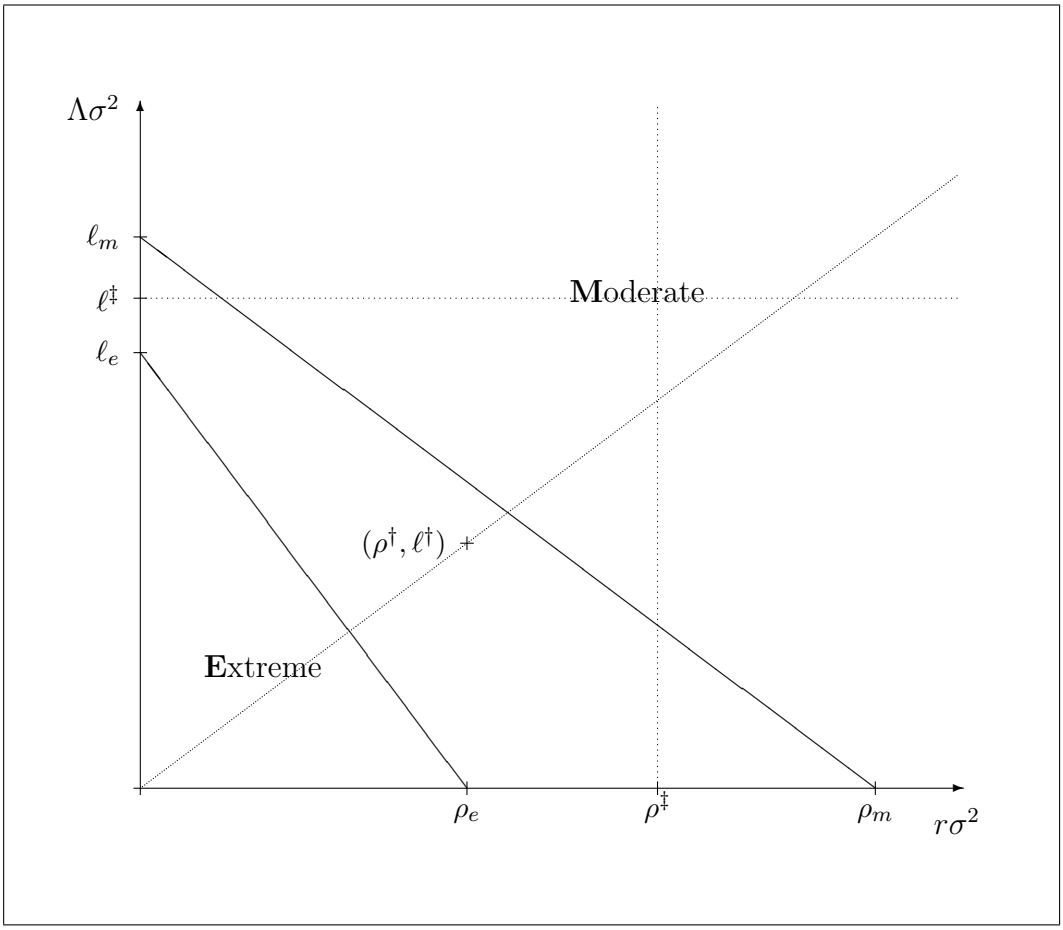


Figure 3.3: Critical parameter values

equation (19). Define $M = \{\mu \geq 0 : v^*[\mu](\hat{\pi}) = \hat{m}\}$ and $\mu^\dagger = \inf M$. By Proposition 4.3, M is non-empty, and $\mu^\dagger$ is finite and positive.

Now fix $\mu_1 \in M$ and $\mu_2 > \mu_1$; we want to show that $\mu_2 \in M$. By Proposition F.2, the second derivative of $v^*[\mu_1]$ is positive on $]0,1[\ -\{\hat{\pi}\}$. Being a *solution* of ODE$[\mu_1]$, $v^*[\mu_1]$ is thus a strict *supersolution* of ODE$[\mu_2]$. (See Appendix G for a definition of supersolution, and compare the proof of Theorem 5.2 below.) By Corollary G.1, therefore, there exists a continuous function $v : [0,1] \to I\!R$ which solves ODE$[\mu_2]$ with $m < v < v^*[\mu_1]$ on $]0,1[\ -\{\hat{\pi}\}$. The uniqueness part of Theorem 4.1 implies $v = v^*[\mu_2]$. So we have $v^*[\mu_2](\hat{\pi}) = \hat{m}$, hence $\mu_2 \in M$.

Combined with a standard continuity argument, this shows that $M = [\mu^\dagger, \infty[$. We can thus set $(\rho^\dagger, \ell^\dagger) = (\mu^\dagger \rho_0, \mu^\dagger \ell_0)$. ∎

We have the following corollary to the above proposition for the limiting cases where either Λ or r is zero. Given σ, we find thresholds for r (when $\Lambda = 0$) and Λ (when $r = 0$) which separate extreme from moderate experimentation as in Figure 3.3.

Corollary 4.2 *When $\Lambda = 0$, there exists a unique real number $\rho^\ddagger > 0$ such that optimal experimentation is extreme if $r < \rho^\ddagger/\sigma^2$, and moderate if $r \geq \rho^\ddagger/\sigma^2$.*

When $r = 0$, there is a unique real number $\ell^\ddagger > 0$ such that optimal experimentation is extreme if $\Lambda < \ell^\ddagger/\sigma^2$, and moderate if $\Lambda \geq \ell^\ddagger/\sigma^2$.

PROOF: Apply Proposition 4.4 to the rays $\mathbb{R}_+ \times \{0\}$ and $\{0\} \times \mathbb{R}_+$, respectively. ■

We mentioned above that constants as in Proposition 4.3 can be calculated explicitly. This yields explicit upper and lower bounds for each of $\rho^\ddagger$ and $\ell^\ddagger$.

Thus, the boundary between the regions of moderate and extreme experimentation links $\rho^\ddagger$ on the ρ-axis with $\ell^\ddagger$ on the ℓ-axis, and lies between the lines $\rho/\rho_e + \ell/\ell_e = 1$ and $\rho/\rho_m + \ell/\ell_m = 1$. Furthermore, given any point $(\rho^\dagger, \ell^\dagger)$ on the boundary, the set $\left\{(\rho, \ell) : \rho \geq \mu\rho^\dagger, \ell = \mu\ell^\dagger, \mu \geq 1\right\}$ can be shown to lie within the region of moderate experimentation (cf. the comparative statics results mentioned in Section 4.5 below).

Note that within the L-shaped region $\left\{(\rho, \ell) : \rho < \rho^\ddagger \text{ or } \ell < \ell^\ddagger\right\}$, that is, below or to the left of the dotted lines in Figure 3.3, there is a potential trade-off between the discount rate and the switching intensities, given a noise level σ: for any $r < \rho^\ddagger/\sigma^2$, moderate experimentation can be avoided if Λ is sufficiently low, and similarly, for any $\Lambda < \ell^\ddagger/\sigma^2$, moderate experimentation can be avoided if r is sufficiently low.

4.3 No State Switching ($r > 0,\ \Lambda = 0$)

The discounted case with $\Lambda = 0$ is simpler since we do not have to make the transformation from u^* to v^*, but can work with u^* itself. Our next result provides a detailed characterisation of the optimal experimentation behaviour.

Proposition 4.5 *Let $\Lambda = 0$.*

In the case of extreme experimentation *($u^*(\hat{\pi}) > \hat{m}$), the optimal policy is continuous except for a single jump between $q_{\max}$ and $q_{\min}$, and selects one of these quantities at $\hat{\pi}$. The process of posterior beliefs, if started from a prior belief in $]0, 1[$, can reach any other point in this interval, and converges almost surely to the true state of demand.*

In the case of moderate experimentation *($u^*(\hat{\pi}) = \hat{m}$), the optimal policy is continuous and selects $\hat{q}$ at $\hat{\pi}$. Given a prior belief $\pi_0 \neq 0,\ \hat{\pi},\ 1$ and the true state of demand k, the process of posterior beliefs is confined to the subinterval $]0, \hat{\pi}[$ or $]\hat{\pi}, 1[$ which contains π_0, and it exhibits the following long-run behaviour: if π_0 lies between $\hat{\pi}$ and k then beliefs converge almost surely to $\hat{\pi}$ or k (each limit having positive probability); otherwise, beliefs converge almost surely to $\hat{\pi}$.*

PROOF: For $u^*(\hat{\pi}) > \hat{m}$, we obtain the optimal policy exactly as in the proof of Proposition 4.1. Since u^* is strictly convex, its graph crosses $\mathcal{R}_c$ only once, so this policy has indeed just one jump.

If $u^*(\hat{\pi}) = \hat{m}$, on the other hand, we have $(u^*)'(\hat{\pi}) = 0$ and $\tau(\pi)(u^*)''(\pi)/2r = \beta(\pi)[u^*(\pi) - m(\pi)]/[u^*(\pi) - \hat{m}] < \beta(\pi)$ on $]0,1[-\{\hat{\pi}\}$. For all $\pi \neq \hat{\pi}$, there are ξ and ζ strictly between $\hat{\pi}$ and π such that $[u^*(\pi) - \hat{m}]/[m(\pi) - \hat{m}] = (u^*)''(\xi)/m''(\zeta)$; as $(u^*)''$ is bounded on $]0,1[-\{\hat{\pi}\}$, so is the quotient on the left-hand side. Therefore, $O(\pi, u^*(\pi))$ extends to a Lipschitz continuous policy function which is optimal by the verification theorem from Section 3.3.

The updating equation (4) shows that the process of posterior beliefs generated by the optimal policy is a supermartingale if the true state is $k = 0$, and a submartingale if $k = 1$. Since the process is bounded, this implies almost sure convergence. The stated long-run behaviour is now established by means of the standard boundary classification for diffusion processes; cf. Karlin and Taylor (1981, Chapter 15, Sections 6–7). ■

Combined with Corollary 4.2, the proposition shows that for sufficiently high discount rates, there is a positive probability that beliefs will settle down at a point where the agent has not learnt the true state. This is a particular case of the general incomplete learning result obtained in the literature on optimal learning in an unchanging environment and referred to after Proposition 4.2. Our setup with continuous time and a one-dimensional state space makes this result particularly stark: if the prior belief lies on the 'wrong' side of $\hat{\pi}$, moderate experimentation will cause beliefs to converge to $\hat{\pi}$ with probability one!

Figure 3.10 shows u^* $(= v^*)$ and q^* for $r = 0.1$ and $\Lambda = 0$. Now that there is no state switching, the (adjusted) value function is much higher and extreme experimentation is optimal over a wider range of beliefs. Three sample paths are shown in Figure 3.11 where the true state is $k_t = 0$: the bold one is for $\pi_0 = 0.25$; the faint one is also for $\pi_0 = 0.25$ but with less 'benign' shocks; the other one is for $\pi_0 = 0.75$, i.e., the 'wrong' side of $\hat{\pi}$. In each example, the process of beliefs converges to the truth.

Contrast this with graphs for a higher discount rate, $r = 0.5$ and $\Lambda = 0$: in Figure 3.12 neither u^* $(= v^*)$ nor q^* is very different from its myopic counterpart; and the sample paths in Figure 3.13 illustrate the possible limits of the process of beliefs in the case of moderate experimentation, depending on the prior belief π_0.

Finally, note that the case of a changing environment is richer in that it allows for a form of moderate experimentation where the monopolist experiments much more on one side of $\hat{\pi}$ than on the other. This asymmetry reflects the fact that state switching introduces mean reversion into the updating equation (2), thereby

destroying the martingale property of beliefs.

4.4 Maximum Experimentation ($r = 0,\ \Lambda = 0$)

For given σ, an agent who uses the catching-up criterion in an unchanging environment clearly has the strongest possible incentive to experiment. Consequently, if extreme experimentation is to occur at all, it must occur for $r = 0$ and $\Lambda = 0$, and we have seen above that this is indeed the case. Further, we know from Appendix E.1 that the (adjusted) value function u^* $(= v^*)$ coincides with the *ex ante* full-information pay-off function $\overline{m}$, given by $\overline{m}(\pi) = (1-\pi)\, m(0) + \pi\, m(1)$. This enables us to derive the optimal policy in a simple closed form, below. Recall the construction of the left, central and right rays in Appendix D, and that q_c denotes the centre of the interval $[q_{\min}, q_{\max}]$.

Proposition 4.6 *Let $r = 0$ and $\Lambda = 0$. Let π_ℓ, π_c and π_r be the beliefs where the graph of the full-information pay-off function $\overline{m}$ intersects the left, central and right rays, respectively, that is,*

$$\begin{aligned} \overline{m}(\pi_\ell) &= \hat{m} - \tfrac{1}{2}\,\Delta\alpha\,[q_{\max} - \hat{q}]\,(\pi_\ell - \hat{\pi}), \\ \overline{m}(\pi_r) &= \hat{m} + \tfrac{1}{2}\,\Delta\alpha\,[\hat{q} - q_{\min}]\,(\pi_r - \hat{\pi}), \end{aligned}$$

and

$$\pi_c = \hat{\pi} + \frac{2}{\Delta\alpha}\,\frac{q_c - \hat{q}}{[q_{\max} - \hat{q}][\hat{q} - q_{\min}]}\,(\overline{m}(\pi_c) - \hat{m}).$$

Then the policy function $\overline{q} : [0,1] \to Q$ defined by

$$\overline{q}(\pi) = \begin{cases} q^m(\pi) + \frac{\overline{m}(\pi) - m(\pi)}{m(\pi) - \hat{m}}\,[q^m(\pi) - \hat{q}] & \text{for } \ 0 \le \pi < \pi_\ell, \ \text{or } \pi_r < \pi \le 1, \\ q_{\max} & \text{for } \ \pi_\ell \le \pi \le \pi_c, \\ q_{\min} & \text{for } \ \pi_c \le \pi \le \pi_r, \end{cases}$$

which is continuous except for a jump at π_c, is optimal.

PROOF: $\overline{q}$ is measurable with $\overline{q}(\pi) - \hat{q}$ bounded away from zero, hence admissible by Proposition A.1. Optimality follows from the undiscounted variant of the verification theorem from Section 3.3 and the fact that $\overline{q}(\pi) \in O(\pi, \overline{m}(\pi))$ for $0 < \pi < 1$. ∎

Note that this result holds independently of the value of the parameter σ. The reason for this is simple. In the absence of state switching, a change in σ amounts to

a mere rescaling of the time axis. As the objective of an agent using the catching-up criterion is invariant to such a rescaling, the optimal policy remains the same.

4.5 Further Findings

Beyond the results reported above, our numerical simulations suggest additional properties of the adjusted value function and the optimal policy when $r > 0$ or $\Lambda > 0$. While Theorem 4.1 establishes strict convexity of the adjusted value function in the moderate experimentation regime, v^* appears to be strictly convex in the extreme experimentation regime as well; see for example Figure 3.4. This implies in particular that the graph of v^* crosses the central ray only once, so extreme experimentation entails just a single jump in the optimal policy function.

Granted strict convexity of v^*, we also have the strict inequality $v^* < \overline{m}$ on $]0,1[$. This inequality reflects the intuitive fact that the incentive to experiment is highest for an infinitely patient agent who operates in an environment which does not change.

Further, our numerical results suggest that v^* is strictly decreasing (on $]0,1[$ for extreme experimentation, on $]0,1[- \{\hat{\pi}\}$ for moderate experimentation) in each of the parameters r, Λ and σ. As a consequence, the extent to which the agent experiments, measured by the distance $|q^*(\pi) - q^m(\pi)|$, is strictly decreasing in each of these parameters as long as $q^*(\pi) \notin \{q_{\min}, q^m(1), \hat{q}, q^m(0), q_{\max}\}$. In the extreme experimentation regime, moreover, the set of beliefs at which $q_{\max}$ or $q_{\min}$ is optimal shrinks in response to an increase in any of the parameters. (The intuition behind these comparative statics has been given at the beginning of Section 4.2.) For the moderate experimentation regime, we can use the techniques of Appendix G to prove analytically that $v^*(\pi)$, and hence $|q^*(\pi) - q^m(\pi)|$, is always strictly decreasing on $]0,1[- \{\hat{\pi}\}$ in r and σ, and strictly decreasing in Λ in the undiscounted case ($r = 0$); cf. the proof of Theorem 5.2 below.

5 No Confounding Belief

We now turn briefly to the simpler case of optimal experimentation when the two demand curves do not intersect in the interior of Q^m, and so there is no longer a belief $\hat{\pi} \in]0,1[$ such that $q^m(\hat{\pi}) = \hat{q}$.[26] That is, we drop the standing assumption made in Section 3.2, and instead make the following

[26] To avoid cumbersome case distinctions, we will ignore the border-line cases $\hat{q} = q^m(0)$ and $\hat{q} = q^m(1)$ in what follows. It is easy to check, though, that the results given below remain valid in these two scenarios.

Assumption *The quantity $\hat{q}$ does* not *lie strictly between $q^m(0)$ and $q^m(1)$; $q^m(0) \neq q^m(1)$.*

We continue to assume without loss of generality that the demand curve in state 1 is steeper than the demand curve in state 0, that is, $\Delta\beta > 0$. As in Section 3.2, this new assumption is sufficient for experimentation to occur at almost all beliefs.

It turns out that under the new assumption, experimentation is *always* moderate and in one direction. That is, the monopolist never chooses any quantities outside Q^m, and q^* always lies on the same side of $\hat{q}$ as q^m, but further away. In other words, the optimal behaviour in the absence of a confounding belief looks exactly like the long-run behaviour in the moderate experimentation regime of the previous section. In particular, the relevant ODE for the adjusted value function will be given by (20), and optimal quantities by (15).

When $r = 0$ and $\Lambda = 0$, the optimal policy can again be derived in closed form:

$$\overline{q}(\pi) = q^m(\pi) + \frac{\overline{m}(\pi) - m(\pi)}{m(\pi) - \hat{m}} \left[q^m(\pi) - \hat{q}\right] ;$$

since $\overline{q}(\pi) = O(\pi, \overline{m}(\pi))$ for $0 < \pi < 1$, optimality of this policy function follows exactly as in the proof of Proposition 4.6.

When $r > 0$ or $\Lambda > 0$, we have the following.

Theorem 5.1 (Moderate Experimentation, One Direction) *Let $r > 0$ or $\Lambda > 0$. Then the adjusted value function v^* is the unique solution of the ODE* (20) *on $]0,1[$ subject to $v^*(\pi) = m(\pi)$ at $\pi = 0,\ 1$ and $v^* > m$ on $]0,1[$; moreover, it is analytic, strictly convex, and satisfies $v^* < \overline{m}$ on $]0,1[$.*

The optimal policy function,

$$q^*(\pi) = q^m(\pi) + \frac{v^*(\pi) - m(\pi)}{m(\pi) - \hat{m}} \left[q^m(\pi) - \hat{q}\right],$$

is analytic, takes values in Q^m only, and satisfies the following inequalities on $]0,1[$: if $\hat{q} < Q^m$ then $q^m < q^* < \overline{q}$, and if $\hat{q} > Q^m$ then $q^m > q^* > \overline{q}$.

Proof: Applying Proposition G.2, we obtain a continuous function $v : [0,1] \to \mathbb{R}$ which solves (20) with $m < v < \overline{m}$ on $]0,1[$. This function is analytic on $]0,1[$ by the Cauchy-Kowalewski theorem.

Using the verification theorem from Section 3.5 (and its undiscounted variant), we see that $v = v^*$ and the policy q^* is optimal. This also establishes the uniqueness part of the theorem. Moreover, we have $(v^*)'' > 0$ on $]0,1[$ by Proposition F.1.

The stated inequalities for q^* follow directly from the fact that $m < v^* < \overline{m}$ on $]0,1[$. The derivative of the function $\overline{q}$ is

$$\overline{q}'(\pi) = -\frac{[q^m(0) - \hat{q}]\,[q^m(1) - \hat{q}]}{[q^m(\pi) - \hat{q}]^2}\,\frac{\hat{p}\,\Delta\beta}{2\,\beta(\pi)^2} = \frac{[q^m(0) - \hat{q}]\,[q^m(1) - \hat{q}]}{[q^m(\pi) - \hat{q}]^2}\,(q^m)'(\pi)\,;$$

as either $\hat{q} < Q^m$ or $\hat{q} > Q^m$, this is either strictly positive or strictly negative throughout and of the same sign as the derivative of q^m. As $\overline{q}$ and q^m coincide at either end of the unit interval, the range of $\overline{q}$ and hence the range of q^* equals Q^m. ∎

Thus, experimentation is indeed in one direction for $\hat{q} \notin Q^m$: if $\hat{q}$ lies to the left of Q^m, then $q^* > q^m$ on $]0,1[$, which means that the agent experiments by increasing quantity; if, on the other hand, $\hat{q}$ lies to the right of Q^m, then $q^* < q^m$ on $]0,1[$, so the agent experiments by decreasing quantity. The intuition behind this quantity expansion or reduction is straightforward: the monopolist deviates from the myopic quantity by moving away from $\hat{q}$ in the (now unambiguously defined) direction of widening spreads between the two possible demand curves, thus making price observations more informative.[27] Note also that the process of posterior beliefs is now always regular since the difference $q^*(\pi) - \hat{q}$, and hence the informativeness of the price signal, is bounded away from zero.

Convexity of v^* turns out to be crucial for the comparative statics of optimal experimentation, to which we turn next.

Theorem 5.2 *Given fixed demand curve parameters α_0, α_1, β_0, β_1 such that $\hat{q} \notin Q^m$ and a fixed $\tilde{\pi}$, the distance $|q^*(\pi) - q^m(\pi)|$ is*

strictly decreasing in r;

strictly decreasing in σ;

strictly decreasing in Λ if $r = 0$

for all $\pi \in]0,1[$.

PROOF: Let $v^*[r,\Lambda,\sigma]$ denote the adjusted value function for any given combination of parameters $r \geq 0$, $\Lambda \geq 0$ and $\sigma > 0$, and let ODE$[r,\Lambda,\sigma]$ be the differential equation (20) for these parameter values.

Consider discount rates $r_1 < r_2$. Since $v^*[r_1,\Lambda,\sigma] > m$ on $]0,1[$, the right-hand side of (20) with $v = v^*[r_1,\Lambda,\sigma]$ is strictly increasing in r at each $\pi \in]0,1[$. Being a *solution* of ODE$[r_1,\Lambda,\sigma]$, $v^*[r_1,\Lambda,\sigma]$ is thus a strict *supersolution* of ODE$[r_2,\Lambda,\sigma]$. (See Appendix G for a definition of supersolution.) As in Proposition G.2, therefore, there exists a

[27] It is less intuitive, though, that experimentation should always be moderate.

continuous function $v : [0,1] \to \mathbb{R}$ which solves ODE$[r_2, \Lambda, \sigma]$ with $m < v < v^*[r_1, \Lambda, \sigma]$ on $]0,1[$. As $v^*[r_1, \Lambda, \sigma] < \overline{m}$ on $]0,1[$, the uniqueness part of Theorem 5.1 implies $v = v^*[r_2, \Lambda, \sigma]$. The comparative statics result with respect to the discount rate now follows from the observation that optimal quantities are increasing in adjusted values.

The other comparative statics results follow in the same way. In fact, since the second derivative of the adjusted value function is strictly positive on $]0,1[$, $v^*[r, \Lambda, \sigma_1]$ is a strict supersolution of ODE$[r, \Lambda, \sigma_2]$ for $\sigma_1 < \sigma_2$, and $v^*[0, \Lambda_1, \sigma]$ is a strict supersolution of ODE$[0, \Lambda_2, \sigma]$ for $\Lambda_1 < \Lambda_2$. ∎

Again, the intuition behind these comparative statics results has already been discussed in Section 4.2.[28]

6 Conclusion

We have studied the behaviour of a monopolist who learns about randomly changing demand by choosing a stream of quantities, observing the prices they generate, and updating his beliefs accordingly. As the action space is continuous, a small amount of information can be obtained at a small opportunity cost. Given the changing environment, therefore, experimentation will occur even in the long run although, as we have seen, the scope of actions may become restricted.

We formulated the problem in continuous time, which lead us via the Bellman equation to an ordinary differential equation for the adjusted value function. The advantages of this approach are three-fold: (a) key properties of the value function and optimal policy can be established analytically, as can some comparative statics results, even though a closed-form solution is generally not obtainable; (b) the sample path properties of beliefs and optimal actions are easy to characterise; (c) it is straightforward to solve the differential equation of interest numerically, enabling us to illustrate the analytical results and suggest further plausible properties of the solution.

Our analysis focused on the more interesting case where the confounding quantity lies between the myopically optimal quantities for the two possible demand curves.

[28] Clearly, the comparative statics with respect to Λ should also pertain when $r > 0$. Our numerical simulations confirm this conjecture, but we have not been able to provide an analytical proof so far. Note that by the same argument as in the proof of Theorem 5.2, it would be sufficient to show that

$$\left[f(\pi) + (\pi - \tilde{\pi})\frac{\Delta\beta}{\beta(\pi)}\right] \frac{v^*(\pi) - m(\pi)}{v^*(\pi) - \hat{m}} + (\pi - \tilde{\pi}) \left[\frac{v^*(\pi) - m(\pi)}{v^*(\pi) - \hat{m}}\right]' > 0$$

on $]0,1[$, which is equivalent to $(v^*)'' > (u^*)''$ on the open unit interval. All the numerical solutions that we have calculated satisfy this condition.

We found two qualitatively different experimentation regimes. For low discount rates and low probabilities of a demand curve switch, optimal experimentation is extreme: the maximal and minimal feasible quantities are chosen a non-negligible fraction of time; the optimal policy function exhibits a jump from one extreme quantity to the other; and the true state is tracked fairly well. For high discount rates or high probabilities of a demand curve switch, on the other hand, experimentation is moderate: the quantities chosen are bounded away from the extremes; the monopolist behaves like a myopic agent at the confounding belief; and he eventually restricts his choices to a subset of the space of feasible quantities.

A transition from one regime to the other in response to a change in the model parameters involves a discontinuous change of optimal policy. This suggests that agents in a changing environment may reduce their investment in information drastically if the frequency of change (or the interest rate) passes a critical threshold. However, there is a region in which a trade-off between interest rates and stability can be exploited: a moderately high interest rate need not trigger sluggish investment provided that the underlying environment is sufficiently stable; conversely, the low-investment effect of a changing environment can be overcome by a sufficiently low interest rate.

In the case where the confounding quantity does not lie between the two myopically optimal quantities, the agent deviates from myopic behaviour by experimenting towards wider spreads between the demand curves, where the price observations are more informative. This experimentation remains moderate in a well-defined sense and is qualitatively the same for all parameter values.

As to the robustness of our results, the existence of a confounding action is obviously necessary for the emergence of two regimes and the moderate experimentation trap, in the same way as it was necessary for the incomplete learning results of the previous literature on unchanging environments. Here, the linearity of demand curves is inessential.

With non-linear demand curves, on the other hand, the monopolist would not necessarily go all the way to the boundaries of the interval of feasible quantities when experimenting at the confounding belief, i.e. when jumping past the confounding quantity. Moreover, this jump could unfold gradually as parameters change.

Replacing continuous time by discrete time, we would have a less clear-cut behaviour of posterior beliefs under moderate experimentation since discrete adjustments could allow the belief to jump back and forth past the confounding belief. However, adjustments towards the long-run average state become stronger and more frequent the further the current belief is away from the long-run average state, so

excursions out of the trap could be expected to be infrequent and short. With shrinking period length, the resulting sample path behaviour would become very close to that in our model.

Finally, some might think that the agent in our model knows an unrealistic amount at the outset – the demand curve parameters, the switching intensities, the noise level. This abstraction has allowed us to focus clearly on the role of discounting, instability and noise in determining optimal behaviour, but it does not drive our results. The crucial condition is the existence of a confounding action and of a posterior belief at which that action is myopically optimal. As long as this condition holds, the emergence of qualitatively different experimentation regimes and incomplete learning are possible.

Appendix

A Admissible Strategies and Policy Functions

We first provide a precise definition of the set $\mathcal{Q}$ of admissible strategies. Assume that the Brownian motion Z and the Markov process k are given on some complete probability space and are both adapted to the filtration $\{F_t\}$. Let $\mathcal{Q}_0$ denote the set of all processes $\mathbf{q} = \{q_t\}$ which take values in Q, the interval of feasible quantities, and are adapted to the aforementioned filtration. Each $\mathbf{q} \in \mathcal{Q}_0$ gives rise to a unique cumulative price process $P^{\mathbf{q}}$. The information contained in prices is summarised by $\{\mathcal{F}_t^{\mathbf{q}}\}$, the filtration generated by $P^{\mathbf{q}}$. A process $\mathbf{q} \in \mathcal{Q}_0$ is an *admissible strategy* if q_t is adapted to the filtration $\{\mathcal{F}_t^{\mathbf{q}}\}$.

Admissible policy functions can now be defined as follows. The function $q : [0,1] \to Q$ is an *admissible policy function* if for any given initial belief π_0, there is a unique strategy $\mathbf{q} \in \mathcal{Q}$ (with associated process of beliefs $\{\pi_t\}$) such that $q_t = q(\pi_t)$ for all t.

The following result provides conditions under which a given policy function is admissible.

Proposition A.1 *A policy function $q : [0,1] \to Q$ is admissible if at least one of the following conditions holds:*

(a) q is Lipschitz continuous;

(b) q is measurable, and there exists a $\delta > 0$ such that $[\Delta\alpha - \Delta\beta\, q(\pi)]^2 > \delta$ for all π.

Proof: Suppose that (a) holds. Then an extension to a standard existence theorem implies that the stochastic differential equation

$$\begin{aligned} d\pi_t &= \Big\{\lambda(\pi_t) + \sigma^{-1}\,\Sigma(\pi_t, q(\pi_t))\,\Big(\alpha_{k_t} - \beta_{k_t} q(\pi_t) \;-\; [\alpha(\pi_t) - \beta(\pi_t) q(\pi_t)]\Big)\Big\}\,dt \\ &\quad + \Sigma(\pi_t, q(\pi_t))\,dZ_t \end{aligned} \tag{A.1}$$

which is obtained from combining (2) and (3) has a unique solution π for any given starting value $\pi_0 \in [0,1]$; cf. Liptser and Shiryayev (1977, p.330).[29] Define the strategy $\mathbf{q}$ by $q_t = q(\pi_t)$ and consider the associated price process $dP_t = (\alpha_{k_t} - \beta_{k_t}\, q_t)\,dt + \sigma\,dZ_t$. Section 2 implies that the corresponding process of beliefs $\pi_t^{\mathbf{q}} = \mathrm{E}[k_t \mid \mathcal{F}_t^{\mathbf{q}}]$ also solves (A.1) with initial value π_0. By the uniqueness part of Liptser and Shiryayev (1977, Theorem

[29] This is in fact a strong solution. A weak solution would be enough for our purposes.

9.2), the processes π and $\pi^{\mathbf{q}}$ coincide, so π is indeed the process of beliefs associated with the strategy $\mathbf{q}$. The policy function q thus generates a unique strategy in $\mathcal{Q}$.

Now suppose (b). Given any initial value π_0, Krylov (1980, Theorem 2.6.1) implies that the stochastic differential equation

$$d\pi_t = \lambda(\pi_t)\,dt + \Sigma(\pi_t, q(\pi_t))\,dZ_t$$

has a weak solution (π, Z^0) with Z^0 a Wiener process. We extend the corresponding filtered probability space in such a way that it supports an independent Markov process $\{k_t\}$ taking values in $\{0,1\}$ with transition probabilities as in Section 1. Consider the bounded process

$$\eta_t = \sigma^{-1}\Big(\alpha_{k_t} - \beta_{k_t} q(\pi_t) \;-\; [\alpha(\pi_t) - \beta(\pi_t) q(\pi_t)]\Big).$$

By Girsanov's theorem, there is a new measure under which

$$Z_t = Z_t^0 - \int_0^t \eta_s\,ds$$

is a Wiener process; cf. Revuz and Yor (1991). In other words, (π, Z) is a weak solution to the stochastic differential equation (A.1). Admissibility of the policy function q is now shown in exactly the same way as in the first part of this proof. ∎

B Some Properties of the Value Function

Consider the value function u^* as defined in (5).

Proposition B.1 *The value function u^* is continuous and convex.*

PROOF: For fixed $\mathbf{q} \in \mathcal{Q}$, $u^{\mathbf{q}}$ is linear in π. Indeed,

$$\begin{aligned} u^{\mathbf{q}}(\pi) \;&=\; \pi\;\mathrm{E}_{k_0=1}\left[\int_0^\infty r\,e^{-r\,t}\;q_t\,[\alpha_{k_t} - \beta_{k_t} q_t]\,dt\right] \\ &\quad +\;(1-\pi)\;\mathrm{E}_{k_0=0}\left[\int_0^\infty r\,e^{-r\,t}\;q_t\,[\alpha_{k_t} - \beta_{k_t} q_t]\,dt\right]. \end{aligned}$$

For $\pi = \eta\,\pi_1 + (1-\eta)\,\pi_2$ with $0 \le \eta \le 1$, we therefore have

$$\begin{aligned} u^{\mathbf{q}}(\pi) \;&=\; \eta\,u^{\mathbf{q}}(\pi_1) + (1-\eta)\,u^{\mathbf{q}}(\pi_2) \\ &\le\; \eta\,u^*(\pi_1) + (1-\eta)\,u^*(\pi_2) \end{aligned}$$

by the definition of the value function. Taking the supremum on the left-hand side proves convexity. A convex function is continuous on the interior of its domain, so we only have to show continuity at $\pi = 0$ and $\pi = 1$. Suppose for example that the value function is not

continuous at $\pi = 0$. Due to convexity, this can only mean $u^*(0) > u^*(0+)$. By definition of the value function, there exists a policy $\mathbf{q} \in \mathcal{Q}$ such that $u^{\mathbf{q}}(0) > u^*(0+)$. But then $u^{\mathbf{q}}(\pi) > u^*(\pi)$ for small $\pi > 0$, which is a contradiction. The right boundary $\pi = 1$ is dealt with in the same way. ∎

Convexity implies the existence of a left-hand derivative D_-u^* on $]0,1]$ and a right-hand derivative D_+u^* on $[0,1[$, both being non-decreasing functions, the former left-continuous, the latter right-continuous, with $D_-u^* \leq D_+u^*$ on their common domain.

Lemma B.1 *The one-sided derivatives D_-u^* and D_+u^* are bounded.*

PROOF: We see from the representation of the pay-off function $u^{\mathbf{q}}$ in the previous proof that there is a constant $K > 0$ such that $|(u^{\mathbf{q}})'(\pi)| \leq K$ for all $\mathbf{q} \in \mathcal{Q}$ and all π. Now, suppose that $(D_-u^*)(\pi_1) < -K$ for some belief $\pi_1 > 0$. Then there is a $\pi_2 < \pi_1$ such that $u^*(\pi_1) - u^*(\pi_2) < -K(\pi_1 - \pi_2)$, i.e., $u^*(\pi_2) > u^*(\pi_1) + K(\pi_1 - \pi_2)$. By definition of the value function, we can find a strategy $\mathbf{q} \in \mathcal{Q}$ with $u^*(\pi_2) \geq u^{\mathbf{q}}(\pi_2) > u^*(\pi_1) + K(\pi_1 - \pi_2)$. But then the linearity of $u^{\mathbf{q}}$ implies $u^{\mathbf{q}}(\pi_1) \geq u^{\mathbf{q}}(\pi_2) - K(\pi_1 - \pi_2) > u^*(\pi_1)$, which is a contradiction. Using a similar argument for the right-hand derivative, we obtain $-K \leq D_-u^* \leq D_+u^* \leq K$ on $]0,1[$. Due to left- and right-continuity, respectively, this also proves that $(D_-u^*)(1)$ and $(D_+u^*)(0)$ are bounded in absolute value by K. ∎

C The Value Function as a Solution of the Bellman Equation

Proposition C.1 *The value function u^* has a continuous first derivative on $[0,1]$, and possesses a locally bounded generalised second derivative $u_2^* \geq 0$ such that*

$$(u^*)'(\pi_2) - (u^*)'(\pi_1) = \int_{\pi_1}^{\pi_2} u_2^*(\pi)\, d\pi \tag{C.1}$$

for all π_1 and π_2. Moreover,

$$\max_{q \in Q} \left\{ \tfrac{1}{2}\, \Sigma^2(\pi, q)\, u_2^*(\pi) + \ \lambda(\pi)\, (u^*)'(\pi) - r\, u^*(\pi) + r\, R(\pi, q) \right\} = 0 \tag{C.2}$$

almost everywhere on $]0,1[$.

PROOF: Krylov (1980, Theorem 6, p.289) implies that u^* has two locally bounded generalised derivatives, u_1^* and u_2^*. By definition, this means that

$$\int_0^1 \phi(\pi)\, u_1^*(\pi)\, d\pi = -\int_0^1 \phi'(\pi)\, u^*(\pi)\, d\pi$$

and

$$\int_0^1 \phi(\pi)\, u_2^*(\pi)\, d\pi = \int_0^1 \phi''(\pi)\, u^*(\pi)\, d\pi$$

for all functions ϕ that are infinitely differentiable and of compact support in $]0,1[$. On the other hand, u^* is convex by Proposition B.1. As its left-hand derivative D_-u^* is left-continuous and non-decreasing, one can define a measure μ on $]0,1[$ via $\mu[\pi_1,\pi_2[= (D_-u^*)(\pi_2) - (D_-u^*)(\pi_1)$. This measure represents the second derivative of u^* in the sense of a distribution:

$$\int_0^1 \phi''(\pi)\, u^*(\pi)\, d\pi = \int_0^1 \phi(\pi)\, d\mu(\pi)$$

for every function ϕ that is infinitely differentiable and of compact support in $]0,1[$. Moreover, this property characterises μ uniquely; cf. Krylov (1980, p.49). Comparing it with the definition of the generalised second derivative u_2^*, we conclude that $d\mu = u_2^*\, d\pi$. In particular,

$$(D_-u^*)(\pi_2) - (D_-u^*)(\pi_1) = \int_{\pi_1}^{\pi_2} u_2^*(\pi)\, d\pi$$

for all $\pi_1, \pi_2 \in]0,1[$. This implies that D_-u^* is continuous, so u^* is continuously differentiable on the open unit interval with $(u^*)' = D_-u^*$. By Proposition B.1, $(u^*)'(\pi)$ has a continuous extension to the whole of $[0,1]$.

As to the last part of the proposition, Krylov (1980, Theorem 6, p.289) implies that

$$\max_{q\in Q} \left\{ \tfrac{1}{2}\, \Sigma^2(\pi,q)\, u_2^*(\pi) + \ \lambda(\pi)\, u_1^*(\pi) - r\, u^*(\pi) + r\, R(\pi,q) \right\} = 0$$

almost everywhere on $]0,1[$. The proof is completed by replacing u_1^* with $(u^*)'$. ∎

The representation (C.1) implies

Corollary C.1 *u^* is almost everywhere twice differentiable, and $(u^*)'' = u_2^*$ almost everywhere. Moreover, u^* is twice continuously differentiable on any open set where u_2^* has a continuous version, i.e., coincides with a continuous function almost everywhere.* ∎

Applying (C.2) with $q = q^m(\pi)$ and dividing through by r, we see immediately that

$$u^*(\pi) - \lambda(\pi)\frac{(u^*)'(\pi)}{r} \geq m(\pi)$$

almost everywhere. By continuity of u^* and $(u^*)'$, we can conclude that this inequality holds in fact on the whole of $[0,1]$. As to the boundary of the unit interval, we have the following result.

Corollary C.2 *The value function satisfies the boundary conditions*

$$u^*(0) - \lambda(0)\frac{(u^*)'(0)}{r} = m(0)\,, \qquad u^*(1) - \lambda(1)\frac{(u^*)'(1)}{r} = m(1)\,.$$

PROOF: We first note that (C.2) implies

$$\tfrac{1}{2}\max_{q\in Q}\Sigma^2(\pi,q)\,u_2^*(\pi) + \lambda(\pi)\,(u^*)'(\pi) - r\,u^*(\pi) + r\,m(\pi) \geq 0$$

and hence

$$u_2^*(\pi) \geq \frac{2r\sigma^2}{\pi^2(1-\pi)^2}\,\frac{u^*(\pi) - \lambda(\pi)(u^*)'(\pi)/r - m(\pi)}{\max_{q\in Q}[\Delta\alpha - \Delta\beta\,q]^2}$$

for almost all π. Now suppose that $u^*(0) - \lambda(0)(u^*)'(0)/r > m(0)$. Using the continuity of $u^*(\pi) - \lambda(\pi)(u^*)'(\pi)/r$ and the inequality just derived, we can find $K > 0$ and $\epsilon > 0$ such that $u_2^*(\pi) \geq K\pi^{-2}$ almost everywhere on $[0,\epsilon]$. But then

$$(u^*)'(\pi) = (u^*)'(\epsilon) - \int_\pi^\epsilon u_2^*(\xi)\,d\xi \leq (u^*)'(\epsilon) - K\int_\pi^\epsilon \frac{d\xi}{\xi^2} = (u^*)'(\epsilon) - K\left[\frac{1}{\pi} - \frac{1}{\epsilon}\right] \longrightarrow -\infty$$

as $\pi \to 0$, which contradicts the boundedness of $(u^*)'$. The boundary condition at $\pi = 1$ follows by the same argument. ∎

The next result is a so-called *verification theorem*, providing sufficient conditions for a given solution of the Bellman equation to be the value function, and for a given policy function to be optimal or ϵ-optimal.

Proposition C.2 *Let u be a once continuously differentiable function on $[0,1]$ with a generalised second derivative $u_2 \geq 0$ on $]0,1[$ such that*

$$u'(\pi_2) - u'(\pi_1) = \int_{\pi_1}^{\pi_2} u_2(\pi)\,d\pi$$

for all π_1 and π_2, and $\pi^2\,(1-\pi)^2\,u_2(\pi) \to 0$ as $\pi \to 0$ and $\pi \to 1$, respectively. If

$$\max_{q\in Q}\left\{\tfrac{1}{2}\,\Sigma^2(\pi,q)\,u_2(\pi) + \lambda(\pi)\,u'(\pi) - r\,u(\pi) + r\,R(\pi,q)\right\} = 0$$

on $]0,1[$, then the following statements hold true:

(a) $u(\pi) \geq u^{\mathbf{q}}(\pi)$ for all $\mathbf{q} \in \mathcal{Q}$ and all π, that is, $u \geq u^*$.

(b) Let $\epsilon > 0$. If $q : [0,1] \to Q$ is an admissible policy function satisfying

$$\tfrac{1}{2}\,\Sigma^2(\pi,q(\pi))\,u_2(\pi) + \lambda(\pi)\,u'(\pi) - r\,u(\pi) + r\,R(\pi,q(\pi)) \geq -\epsilon\,r \qquad \text{(C.3)}$$

for all π, then the strategy $\mathbf{q}_\pi$ obtained by following this policy from any given initial belief π is ϵ-optimal, i.e., $u^{\mathbf{q}_\pi}(\pi) \geq u(\pi) - \epsilon$. In particular, $u^* \geq u - \epsilon$.

(c) *If there is an admissible policy function as in (b) for any $\epsilon > 0$, then u is the value function: $u = u^*$.*

(d) *If $q^* : [0,1] \to Q$ is an admissible policy function such that for every π, the quantity $q^*(\pi)$ attains the supremum in* (C.2)*, then this policy function is optimal. For any π,*

$$u(\pi) = u^*(\pi) = \max_{\mathbf{q} \in \mathcal{Q}} u^{\mathbf{q}}(\pi) = u^{\mathbf{q}^*_\pi}(\pi)$$

where $\mathbf{q}^*_\pi$ is the strategy obtained by following this policy from the initial belief π.

PROOF: Let the initial belief be $\pi_0 = \pi$. For an arbitrary strategy $\mathbf{q} \in \mathcal{Q}$ consider the stochastic process $M^{\mathbf{q}}$ given by

$$M^{\mathbf{q}}_T = \int_0^T r\, e^{-r\,t}\ R(\pi_t, q_t)\, dt \ +\ e^{-r\,T}\ u(\pi_T)\,.$$

By a generalisation of Itô's lemma,

$$\begin{aligned} M^{\mathbf{q}}_T \ &= \ M^{\mathbf{q}}_0 \\ &\quad + \int_0^T e^{-r\,t} \Big\{ \tfrac{1}{2}\,\Sigma^2(\pi_t, q_t) u_2(\pi_t) + \lambda(\pi_t)\, u'(\pi_t) - r\, u(\pi_t) + r\, R(\pi_t, q_t) \Big\}\, dt \\ &\quad + \sigma^{-1} \int_0^T e^{-r\,t} \pi_t(1-\pi_t)(\Delta\alpha - \Delta\beta\, q_t)\, dZ^{\mathbf{q}}_t\,; \end{aligned}$$

cf. Rogers and Williams (1987, Lemma IV.45.9, p.105). Now, (C.2) implies that the expression under the first integral is non-positive, so $M^{\mathbf{q}}$ is a supermartingale. In other words, $\mathrm{E}_\pi[M^{\mathbf{q}}_T] \le M^{\mathbf{q}}_0$ or

$$u(\pi) \ge \mathrm{E}_\pi \left[\int_0^T r\, e^{-r\,t}\ R(\pi_t, q_t)\, dt \right] + e^{-r\,T}\ \mathrm{E}_\pi[u(\pi_T)]\,.$$

Letting T go to infinity, we see that the first term on the right-hand side becomes $u^{\mathbf{q}}(\pi)$, while the second term tends to zero. This proves part (a). Next, let $\epsilon \ge 0$, and consider a policy function $q : [0,1] \to Q$ satisfying (C.3) on the whole of its domain. If $\mathbf{q}$ is the strategy obtained by following this policy from the initial belief π, then

$$\mathrm{E}_\pi[M^{\mathbf{q}}_T] \ge M^{\mathbf{q}}_0 - \epsilon \int_0^T r\, e^{-r\,t}\, dt\,.$$

Letting $T \to \infty$ yields $u^{\mathbf{q}}(\pi) \ge u(\pi) - \epsilon$. Parts (b), (c) and (d) follow immediately. ∎

D Analysing the Bellman Equation

In Section 3.3 of the main text we initially rewrote the Bellman equation in the form

$$v = \max_{q \in Q} \left\{ \tau(\pi)\, s\, [q - \hat{q}]^2 + R(\pi, q) \right\}$$

where the variable v is standing in for $u(\pi) - \lambda(\pi)u'(\pi)/r$ and s is representing $u''(\pi)/2r$. The first task here is to show that this problem can be reformulated as

$$\tau(\pi)\, s = \min_{q \in Q - \{\hat{q}\}} \frac{v - R(\pi, q)}{[q - \hat{q}]^2}$$

for triplets (π, v, s) with $s \geq 0$ and (π, v) lying in the set

$$\mathcal{A} = \{(\pi, v) \in \,]0, 1[\, \times \mathbb{R} : \; v \geq m(\pi) \text{ and } v > \hat{m}\}\,.$$

(As noted in the main text, the condition $v > \hat{m}$ only bites if $\hat{q}$ lies in the interior of Q^m, in which case it rules out exactly the point $(\hat{\pi}, \hat{m})$, which in turn excludes the possibility of $\hat{q}$ being optimal.)

To derive the reformulation, define the functions

$$B[\pi, v, s, q] = \tau(\pi)\, s\, [q - \hat{q}]^2 + R(\pi, q) - v$$

and

$$B^*[\pi, v, s] = \max_{q \in Q} B[\pi, v, s, q]\,,$$

and rewrite the Bellman equation as

$$B^*[\pi, v, s] = 0\,.$$

Then, for all triplets (π, v, s) with $s \geq 0$ and $(\pi, v) \in \mathcal{A}$, the equation $B^*[\pi, v, s] = 0$ is equivalent to $\max_{q \in Q - \{\hat{q}\}} B[\pi, v, s, q] = 0$. Now, we have $[q - \hat{q}]^2 > 0$ on $Q - \{\hat{q}\}$, so $\max_{q \in Q - \{\hat{q}\}} B[\pi, v, s, q] = 0$ if and only if

$$\max_{q \in Q - \{\hat{q}\}} \frac{B[\pi, v, s, q]}{[q - \hat{q}]^2} = 0\,,$$

which in turn is equivalent to

$$\tau(\pi)\, s = \min_{q \in Q - \{\hat{q}\}} \frac{v - R(\pi, q)}{[q - \hat{q}]^2}\,. \tag{D.1}$$

Moreover, a quantity $q^* \in Q - \{\hat{q}\}$ satisfies $B[\pi, v, s, q^*] = B^*[\pi, v, s] = 0$ if and only if (D.1) holds and q^* minimises $[v - R(\pi, q)]/[q - \hat{q}]^2$. Thus, the original problem and its reformulation are equivalent on $\mathcal{A}$.

Next we shall prove the claim made in Section 3.4 that, when $\hat{q}$ lies in the interior of the interval Q^m, the area $\mathcal{A}$ can be subdivided into four regions by the following rays emanating from $(\hat{\pi}, \hat{m})$:

$$\begin{aligned}
\mathcal{R}_\ell &= \left\{(\pi, v) \in \mathcal{A} : v = \hat{m} - \tfrac{1}{2}\,\Delta\alpha\,[q_{\max} - \hat{q}]\,(\pi - \hat{\pi})\right\}, \\
\mathcal{R}_r &= \left\{(\pi, v) \in \mathcal{A} : v = \hat{m} + \tfrac{1}{2}\,\Delta\alpha\,[\hat{q} - q_{\min}]\,(\pi - \hat{\pi})\right\}, \\
\mathcal{R}_c &= \left\{(\pi, v) \in \mathcal{A} : \pi = \hat{\pi} + \frac{2}{\Delta\alpha}\,\frac{q_c - \hat{q}}{[q_{\max} - \hat{q}]\,[\hat{q} - q_{\min}]}\,(v - \hat{m})\right\},
\end{aligned}$$

the regions being associated with cases in which the above optimisation problems have interior or corner solutions.

Before proceeding, we use equation (9) to replace $R(\pi, q)$ in the optimisation problems to obtain

$$v - m(\pi) = \max_{q \in Q} \left\{\tau(\pi)\, s\,[q - \hat{q}]^2 - \beta(\pi)\,[q - q^m(\pi)]^2\right\} \tag{D.2}$$

and its reformulation

$$\tau(\pi)\, s = \min_{q \in Q - \{\hat{q}\}} \left\{\frac{\beta(\pi)\;[q - q^m(\pi)]^2 + v(\pi) - m(\pi)}{[q - \hat{q}]^2}\right\}. \tag{D.3}$$

We first provide a preliminary lemma showing the regions of $\mathcal{A}$ where the appropriate second order condition for the above equivalent problems is satisfied. Note that the relationships $m(\pi) = \hat{m} + \beta(\pi)\;[q^m(\pi) - \hat{q}]^2$ and $\beta(\pi)\;[q^m(\pi) - \hat{q}] = -\frac{1}{2}\,\Delta\alpha\,(\pi - \hat{\pi})$ are used in a number of the algebraic manipulations.

Lemma D.1 *Let $\hat{q}$ lie in the interior of Q^m. For $(\pi, v, s) \in \mathcal{A} \times I\!R_+$, the second order condition for the minimisation problem in* (D.3) *is satisfied if and only if (π, v) lies below $\mathcal{R}_{2\ell}$ or below $\mathcal{R}_{2r}$, where*

$$\mathcal{R}_{2\ell} = \{(\pi, v) \in \mathcal{A} : v = \hat{m} - \Delta\alpha\,[q_{\max} - \hat{q}]\,(\pi - \hat{\pi})\}$$

and

$$\mathcal{R}_{2r} = \{(\pi, v) \in \mathcal{A} : v = \hat{m} + \Delta\alpha\,[\hat{q} - q_{\min}]\,(\pi - \hat{\pi})\}\,.$$

PROOF: The second order condition for the minimisation problem in (D.3) is satisfied wherever the second order condition for the maximisation problem in (D.2) is satisfied, and it is clear from (D.2) that the latter holds if and only if $\tau(\pi)\, s - \beta(\pi) < 0$.

Using the inequality $v < \hat{m} - \Delta\alpha\,[q_{\max} - \hat{q}]\,(\pi - \hat{\pi})$ in (D.2) leads, after some manipulation, to

$$\max_{q \in Q} \Big\{ [\tau(\pi)\, s - \beta(\pi)]\,[q - \hat{q}]^2 - 2\beta(\pi)\,[q^m(\pi) - \hat{q}]\,[q_{\max} - q] \Big\} < 0.$$

Evaluating the maximand at $q = q_{\max}$ gives us $[\tau(\pi)\, s - \beta(\pi)]\,[q_{\max} - \hat{q}]^2 < 0$ and so $\tau(\pi)\, s - \beta(\pi) < 0$.

On the other hand, using the inequality $v \geq \hat{m} - \Delta\alpha\,[q_{\max} - \hat{q}]\,(\pi - \hat{\pi})$ for $\pi \leq \hat{\pi}$ we arrive at

$$\max_{q \in Q} \Big\{ [\tau(\pi)\, s - \beta(\pi)]\,[q - \hat{q}]^2 - 2\beta(\pi)\,[q^m(\pi) - \hat{q}]\,[q_{\max} - q] \Big\} \geq 0.$$

The term $2\beta(\pi)\,[q^m(\pi) - \hat{q}]\,[q_{\max} - q]$ is non-negative for $\pi \leq \hat{\pi}$ so in this case $\tau(\pi)\, s - \beta(\pi) \geq 0$.

This proves the assertion concerning $\mathcal{R}_{2\ell}$. The case for $\mathcal{R}_{2r}$ is proved in the same way simply by replacing $q_{\max}$ by $q_{\min}$ and $\pi \leq \hat{\pi}$ by $\pi \geq \hat{\pi}$. ∎

The next lemma shows that the above optimisation problems have an interior solution if and only if (π, v) lies below $\mathcal{R}_\ell$ or below $\mathcal{R}_r$.

Lemma D.2 *Let $\hat{q}$ lie in the interior of Q^m. For $(\pi, v, s) \in \mathcal{A} \times I\!R_+$, the minimisation problem in* (D.3) *has an interior solution if and only if $(\pi, v) \in \mathcal{A}_{\text{int},\ell} \cup \mathcal{A}_{\text{int},r}$, where*

$$\mathcal{A}_{\text{int},\ell} = \Big\{ (\pi, v) \in \mathcal{A} : v < \hat{m} - \tfrac{1}{2}\,\Delta\alpha\,[q_{\max} - \hat{q}]\,(\pi - \hat{\pi}) \Big\}$$

and

$$\mathcal{A}_{\text{int},r} = \Big\{ (\pi, v) \in \mathcal{A} : v < \hat{m} + \tfrac{1}{2}\,\Delta\alpha\,[\hat{q} - q_{\min}]\,(\pi - \hat{\pi}) \Big\}.$$

Moreover, the minimising quantity is given by

$$q^m(\pi) + \frac{v(\pi) - m(\pi)}{m(\pi) - \hat{m}}\,[q^m(\pi) - \hat{q}]$$

and the corresponding minimum is

$$\beta(\pi) \frac{v(\pi) - m(\pi)}{v(\pi) - \hat{m}}\,.$$

Proof: In light of the preceding lemma, the minimisation problem in (D.3) has an interior solution if and only if the first order condition is satisfied when (π, v) lies below $\mathcal{R}_{2\ell}$ or below $\mathcal{R}_{2r}$. Note that $\mathcal{A}_{\text{int},\ell}$ lies below $\mathcal{R}_{2\ell}$ and $\mathcal{A}_{\text{int},r}$ lies below $\mathcal{R}_{2r}$.

The first order condition for the minimisation problem in (D.3) is satisfied by the

quantity

$$q = q^m(\pi) + \frac{v - m(\pi)}{\beta(\pi)\ [q^m(\pi) - \hat{q}]}\,. \tag{D.4}$$

In the borderline cases, this first order condition holds for $q = q_{\max}$ or $q_{\min}$. With $q^\dagger$ denoting either $q_{\max}$ or $q_{\min}$ this can be characterised by

$$q^\dagger = q^m(\pi) + \frac{v - m(\pi)}{\beta(\pi)\ [q^m(\pi) - \hat{q}]}$$

rearranged to give

$$v = m(\pi) + \beta(\pi)\ [q^m(\pi) - \hat{q}]\ [q_{\max} - q^m(\pi)] > m(\pi) \ \text{ iff } \pi < \hat{\pi}$$

and

$$v = m(\pi) - \beta(\pi)\ [q^m(\pi) - \hat{q}]\ [q^m(\pi) - q_{\min}] > m(\pi) \ \text{ iff } \pi > \hat{\pi}$$

where the inequalities are obvious if one notes that $q^m(\pi) - \hat{q}$ has the opposite sign to $\pi - \hat{\pi}$.[30] We have the alternative formulations

$$v = \hat{m} - \tfrac{1}{2}\,\Delta\alpha\ [q_{\max} - \hat{q}]\ (\pi - \hat{\pi})$$

and

$$v = \hat{m} + \tfrac{1}{2}\,\Delta\alpha\ [\hat{q} - q_{\min}]\ (\pi - \hat{\pi})\,.$$

Now, the first order condition holds for some $q \in]q_{\min}, q_{\max}[$ if and only if

$$q_{\min} < q^m(\pi) + \frac{v - m(\pi)}{\beta(\pi)\ [q^m(\pi) - \hat{q}]} < q_{\max}\,;$$

the first inequality is equivalent to $(\pi, v) \in \mathcal{A}_{\text{int},r}$, and the second to $(\pi, v) \in \mathcal{A}_{\text{int},\ell}$.

The expression given for the minimising quantity is simply a manipulation of the right-hand side of (D.4), which when substituted into (D.3) yields the expression for the corresponding minimum. ∎

Finally, we show that the optimisation problems have the unique corner solution $q_{\max}$ if (π, v) lies on or above $\mathcal{R}_\ell$ but to the left of $\mathcal{R}_c$, and the unique corner solution $q_{\min}$ if (π, v) lies on or above $\mathcal{R}_r$ but to the right of $\mathcal{R}_c$; for $(\pi, v) \in \mathcal{R}_c$ both corner solutions are optimal.

Lemma D.3 *Let $\hat{q}$ lie in the interior of Q^m. For $(\pi, v, s) \in \mathcal{A} \times \mathbb{R}_+$, the minimisation problem in* (D.3) *has the corner solution $q_{\max}$ if and only if $(\pi, v) \in \mathcal{A}_{\max} \cup \mathcal{R}_c$,*

[30] Using these two inequalities, it is easy to see that $\mathcal{R}_\ell$ cuts the axis $\pi = 0$ above $m(0)$, and that $\mathcal{R}_r$ cuts the vertical line $\pi = 1$ above $m(1)$.

and the corner solution $q_{\min}$ *if and only if* $(\pi, v) \in \mathcal{A}_{\min} \cup \mathcal{R}_c$*, where*

$$\mathcal{A}_{\max} = \Bigg\{ (\pi, v) \in \mathcal{A} : v \geq \hat{m} - \tfrac{1}{2}\,\Delta\alpha\,[q_{\max} - \hat{q}]\,(\pi - \hat{\pi}) \\ \textit{and } \pi < \hat{\pi} + \frac{2}{\Delta\alpha}\,\frac{q_c - \hat{q}}{[q_{\max} - \hat{q}]\,[\hat{q} - q_{\min}]}\,(v - \hat{m}) \Bigg\}$$

and

$$\mathcal{A}_{\min} = \Bigg\{ (\pi, v) \in \mathcal{A} : v \geq \hat{m} + \tfrac{1}{2}\,\Delta\alpha\,[\hat{q} - q_{\min}]\,(\pi - \hat{\pi}) \\ \textit{and } \pi > \hat{\pi} + \frac{2}{\Delta\alpha}\,\frac{q_c - \hat{q}}{[q_{\max} - \hat{q}]\,[\hat{q} - q_{\min}]}\,(v - \hat{m}) \Bigg\}.$$

PROOF: In light of the previous lemma, we know that corner solutions will prevail in the regions under consideration. Also, it is easy to see from the alternative parameterisation of the central ray, namely

$$v = \hat{m} + \tfrac{1}{2}\,\Delta\alpha \frac{[q_{\max} - \hat{q}]\,[\hat{q} - q_{\min}]}{q_c - \hat{q}}\,(\pi - \hat{\pi})$$

for $q_c \neq \hat{q}$, where $q_c = \frac{1}{2}\,(q_{\min} + q_{\max})$, that $\mathcal{R}_c$ lies between $\mathcal{R}_\ell$ and $\mathcal{R}_r$.

In this region, the borderline case arises when $q_{\max}$ and $q_{\min}$ are both optimal and give the same value of $\tau(\pi)\,s$ in (D.3). This is the case if and only if

$$\frac{\beta(\pi)\;[q_{\max} - q^m(\pi)]^2 + v - m(\pi)}{[q_{\max} - \hat{q}]^2} = \frac{\beta(\pi)\;[q_{\min} - q^m(\pi)]^2 + v - m(\pi)}{[q_{\min} - \hat{q}]^2}$$

and simplification leads to

$$\pi = \hat{\pi} + \frac{2}{\Delta\alpha}\,\frac{q_c - \hat{q}}{[q_{\max} - \hat{q}][\hat{q} - q_{\min}]}\,(v - \hat{m})\,.$$

Thus both extreme quantities are optimal for $(\pi, v) \in \mathcal{R}_c$.

It follows immediately that $q_{\max}$ is uniquely optimal for $(\pi, v) \in \mathcal{A}_{\max}$, to the left of $\mathcal{R}_c$, and $q_{\min}$ is uniquely optimal for $(\pi, v) \in \mathcal{A}_{\min}$, to the right of $\mathcal{R}_c$. ∎

E The Undiscounted Case

In the absence of discounting, the monopolist uses the *catching-up criterion* to choose amongst admissible strategies: given a prior belief π, he looks for a strategy $\mathbf{q}^* \in \mathcal{Q}$ which, in the long run, does at least as well as any other strategy in the

sense that $\liminf_{T\to\infty} \mathrm{E}_\pi[R_T^{\mathbf{q}^*} - R_T^{\mathbf{q}}] \geq 0$ for all $\mathbf{q} \in \mathcal{Q}$, where

$$R_T^{\mathbf{q}} = \int_0^T q_t \left[(\alpha_{k_t} - \beta_{k_t} q_t)\, dt \;+\; \sigma dZ_t\right]$$

is the process of cumulative revenues. The agent achieves this goal by maximising the *transient pay-off*, that is, total expected revenue net of the highest possible long-run average pay-off. Indeed, it can be shown that a strategy which achieves the maximum transient pay-off is catching-up optimal.

E.1 No State Switching ($r = 0,\ \Lambda = 0$)

Let $\Lambda = 0$, so the state of demand is fixed over time. Then, the monopolist can achieve a long-run average pay-off arbitrarily close to the full-information pay-off, that is, $m(0)$ if the true state is $k = 0$, and $m(1)$ if the true state is $k = 1$. In fact, it suffices to follow any admissible policy which coincides with the myopic policy q^m for beliefs close to 0 and 1, and is bounded away from the confounding quantity $\hat{q}$ in case the latter lies in the interior of Q^m, the range of the myopic policy. By the martingale convergence theorem and the standard boundary classification for diffusion processes, beliefs will converge to the truth with probability one,[31] and the quantity chosen will approach the quantity which is optimal for the true demand. Given the initial belief π, the agent's objective is therefore to maximise the transient pay-off

$$u^{\mathbf{q}}(\pi) = \mathrm{E}_\pi \left[\int_0^\infty [R(k, q_t) - m(k)]\, dt\right]$$

where k is the unknown state of demand. By the law of iterated expectations,

$$u^{\mathbf{q}}(\pi) = \mathrm{E}_\pi \left[\int_0^\infty [R(\pi_t, q_t) - \overline{m}(\pi_t)]\, dt\right]$$

where

$$\overline{m}(\pi) = (1 - \pi)\, m(0) + \pi\, m(1)$$

is the *ex ante full-information pay-off*.

Standard results imply that the value function $u^*(\pi) = \sup_{\mathbf{q}\in\mathcal{Q}} u^{\mathbf{q}}(\pi)$ solves the Bellman equation

$$\max_{q\in Q} \left\{\tfrac{1}{2}\, \Sigma^2(\pi, q)\, u''(\pi) - \overline{m}(\pi) + R(\pi, q)\right\} = 0 \tag{E.1}$$

[31] We are assuming here that the agent does not assign prior probability zero to the true state. See Karlin and Taylor (1981) for the classification of boundary points.

subject to the boundary conditions $u(0) = u(1) = 0$. Moreover, if a function u solves (E.1) with these boundary conditions, and there is an admissible policy function $q : [0, 1] \to Q$ such that

$$q(\pi) \in \arg\max_{q \in Q} \left\{ \tfrac{1}{2}\, \Sigma^2(\pi, q)\, u''(\pi) - \overline{m}(\pi) + R(\pi, q) \right\} \tag{E.2}$$

for all π, then $u = u^*$ and the given policy is optimal.

In view of the results of Section 3.3, property (E.2) is equivalent to $q(\pi) \in O(\pi, \overline{m}(\pi))$. Moreover, the affine function $\overline{m}$ trivially solves the ODE (19) for $r = 0$ and $\Lambda = 0$ with $\overline{m}(0) = m(0)$ and $\overline{m}(1) = m(1)$ as boundary values. We can therefore simply define the adjusted value function as $v^* = \overline{m}$.

E.2 State Switching ($r = 0,\ \Lambda > 0$)

We now assume $\Lambda > 0$. Let θ^* be the highest long-run average pay-off achievable with a strategy $\mathbf{q} \in \mathcal{Q}$. According to the introductory remarks to this section, the monopolist's objective is then to maximise

$$u^{\mathbf{q}}(\pi) = \mathrm{E}_\pi \left[\int_0^\infty [R(k_t, q_t) - \theta^*]\, dt \right] = \mathrm{E}_\pi \left[\int_0^\infty [R(\pi_t, q_t) - \theta^*]\, dt \right],$$

the transient pay-off as measured against the benchmark θ^*.

It can be shown that θ^* and the value function $u^*(\pi) = \sup_{\mathbf{q} \in \mathcal{Q}} u^{\mathbf{q}}(\pi)$ solve the Bellman equation

$$\max_{q \in Q} \left\{ \tfrac{1}{2}\, \Sigma^2(\pi, q)\, u''(\pi) + \lambda(\pi)\, u'(\pi) - \theta + R(\pi, q) \right\} = 0 \tag{E.3}$$

almost everywhere subject to the boundary conditions $\theta - \lambda(0)\, u'(0) = m(0)$ and $\theta - \lambda(1)\, u'(1) = m(1)$.[32] Conversely, if a real number θ and a function u solve (E.3) with the stated boundary conditions, and there is an admissible policy function $q : [0, 1] \to Q$ such that

$$q(\pi) \in \arg\max_{q \in Q} \left\{ \tfrac{1}{2}\, \Sigma^2(\pi, q)\, u''(\pi) + \lambda(\pi)\, u'(\pi) - \theta + R(\pi, q) \right\} \tag{E.4}$$

for all π, then $\theta = \theta^*$, u is the value function up to a constant of integration, and the given policy is optimal.

[32] Moreover, u^* possesses the same regularity properties as the value function in the discounted case; arguments similar to those given in Appendices B and C apply. The discussion in Sections 3.1 and 3.2 regarding conditions for experimentation and the confounding quantity carries over as well.

Given the above boundary conditions and the fact that (E.4) is equivalent to $q(\pi) \in O(\pi, \theta - \lambda(\pi)\, u'(\pi))$, we define the adjusted value function by $v^*(\pi) = \theta^* - \lambda(\pi)\,(u^*)'(\pi)$. As the analysis in Sections 3.3 and 3.4 remains valid for $r = 0$ with v now standing for $\theta - \lambda(\pi)u'(\pi)$ and $u''(\pi)/2r$ being replaced by $u''(\pi)/2$, we can argue as in Section 3.5, and show that this v^* is indeed a solution (with the caveats stated there) of the ODE (19) for $r = 0$ and $\Lambda > 0$.

F Convexity of the Adjusted Value Function

Fix $\tilde{\pi} \in]0,1[$, $\Lambda > 0$ and $r \geq 0$. For $k > 0$, let $w_k :]0,1[- \{\tilde{\pi}\} \to I\!R$ be the function defined implicitly by the equation

$$G(\pi, w_k(\pi)) = k\,\tau(\pi)\,|\pi - \tilde{\pi}|^{-2-r/\Lambda}$$

with G as in Section 3.3. Then we have the following facts:

- $w_k \geq m$, with equality at $\pi = 0$ and 1, and a strict inequality everywhere else;
- w_k has a pole at $\pi = \tilde{\pi}$;
- $w_k''(\pi) > 0$ unless $(\pi, w_k(\pi)) \in \mathcal{R}_c$ (see Lemma F.2 below);
- $r\,G(\pi, w_k(\pi)) + \Lambda \left\{ f(\pi)\,G(\pi, w_k(\pi)) + (\pi - \tilde{\pi}) \frac{d}{d\pi} G(\pi, w_k(\pi)) \right\} = 0$ for all $\pi \neq \tilde{\pi}$.

This last property explains our interest in the family of functions w_k, and leads to the following characterisation for the curvature of v^* on the set $\{\pi : v^*(\pi) > m(\pi)\}$.[33]

Lemma F.1 *Let π be such that $v^*(\pi) > m(\pi)$ and $(\pi, v^*(\pi)) \notin \mathcal{R}_c$. If $\pi = \tilde{\pi}$, then $(v^*)''(\pi) > 0$. If $\pi \neq \tilde{\pi}$, let w be that function w_k which coincides with v^* at π; then $(v^*)''(\pi) > 0$ if and only if either $\pi < \tilde{\pi}$ and $(v^*)'(\pi) < w'(\pi)$, or $\pi > \tilde{\pi}$ and $(v^*)'(\pi) > w'(\pi)$.*

Proof: The case $\pi = \tilde{\pi}$ is trivial. At $\pi \neq \tilde{\pi}$, v^* solves the ODE (19), so the fourth property above implies

$$\tau(\pi) \frac{(v^*)''(\pi)}{2} = \Lambda\,(\pi - \tilde{\pi}) \left\{ \frac{d}{d\pi} G(\pi, v^*(\pi)) - \frac{d}{d\pi} G(\pi, w(\pi)) \right\}.$$

The lemma follows since $\frac{d}{d\pi} G(\pi, v(\pi))$ is strictly increasing in $v'(\pi)$. ∎

[33] Note that if $v^*(\pi) = m(\pi)$ with $\pi \neq \hat{\pi}$, then trivially $(v^*)''(\pi) > 0$ since $m''(\pi) > 0$.

We shall use this to prove that the adjusted value function is strictly convex whenever experimentation is moderate.

Proposition F.1 *If $\hat{q}$ is not in the interior of Q^m, then $(v^*)'' > 0$ on $]0,1[$.*

PROOF: Suppose that $(v^*)''(\pi^\dagger) \leq 0$ with $0 < \pi^\dagger < 1$; without loss of generality, we assume that $\pi^\dagger < \tilde{\pi}$. Let $k^\dagger > 0$ be such that the function $w^\dagger = w_{k^\dagger}$ coincides with v^* at $\pi^\dagger$. By the above lemma, $(v^*)'(\pi^\dagger) \geq (w^\dagger)'(\pi^\dagger)$. Now, the strict convexity of the functions w_k implies that $(v^*)'(\pi) > w_k'(\pi)$ for $\pi < \pi^\dagger$ sufficiently close to $\pi^\dagger$ and $k < k^\dagger$ sufficiently close to $k^\dagger$. So $(v^*)''$ is strictly negative immediately to the left of $\pi^\dagger$. On the other hand, v^* cannot be strictly concave on the whole of $]0,\pi^\dagger[$ since this would imply $(v^*)' > w'$ on $]0,\pi^\dagger[$ and hence $v^*(0) < w^\dagger(0) = m(0)$. Therefore, there must be positive $\pi < \pi^\dagger$ such that $(v^*)''(\pi) = 0$ again; let $\pi^\ddagger$ be the biggest such π, and $w^\ddagger$ that function w_k which is tangent to v^* at $\pi^\ddagger$. On $[\pi^\ddagger,\pi^\dagger]$, $w^\ddagger$ and $w^\dagger$ are strictly convex, while v^* is strictly concave. This implies $w^\ddagger(\pi^\ddagger) < w^\dagger(\pi^\ddagger)$ and $w^\ddagger(\pi^\dagger) > w^\dagger(\pi^\dagger)$, so $w^\ddagger$ and $w^\dagger$ must intersect somewhere on $]\pi^\ddagger,\pi^\dagger[$ – a contradiction. The same argument can be used to the right of $\tilde{\pi}$. ∎

Proposition F.2 *Suppose that $\hat{q}$ lies in the interior of Q^m and $v^*(\hat{\pi}) = \hat{m}$. Then v^* is strictly convex with $(v^*)'' > 0$ on $]0,1[- \{\hat{\pi}\}$, and at most one of the one-sided derivatives $(D_-v^*)(\hat{\pi})$ and $(D_+v^*)(\hat{\pi})$ can be different from zero. In fact, if $\hat{\pi} < \tilde{\pi}$, then $(D_+v^*)(\hat{\pi}) = 0$ and $(D_-v^*)(\hat{\pi}) \leq 0$; and if $\hat{\pi} > \tilde{\pi}$, then $(D_-v^*)(\hat{\pi}) = 0$ and $(D_+v^*)(\hat{\pi}) \geq 0$. If v^* is differentiable at $\hat{\pi}$, then the ratio $[v^*(\pi) - \hat{m}]/[m(\pi) - \hat{m}]$ converges to a finite limit as $\pi \to \hat{\pi}$. If v^* has a kink at $\hat{\pi}$, then $[v^*(\pi) - \hat{m}]/[m(\pi) - \hat{m}]$ converges to a finite limit as π approaches $\hat{\pi}$ from the direction of $\tilde{\pi}$.*

PROOF: We first convince ourselves that we can assume $q_c = \hat{q}$ without loss of generality. In fact, suppose we have shown the stated properties for the adjusted value function v^* in this particular case. Then the boundary conditions and strict convexity of v^* imply that $v^* < \overline{m}_\ell$ on $]0,\hat{\pi}[$ and $v^* < \overline{m}_r$ on $]\hat{\pi},1[$ with $\overline{m}_\ell$ and $\overline{m}_r$ being the functions whose graphs are the straight lines joining the point $(\hat{\pi},\hat{m})$ with the points $(0,m(0))$ and $(1,m(1))$, respectively. In particular, the graph of v^* lies entirely in the closure of $\mathcal{A}_{\text{int},\ell} \cup \mathcal{A}_{\text{int},r}$. Arguing exactly as in the proof of Proposition 4.2, we see that v^* is the adjusted value function for *any* $q_{\max}$ and $q_{\min}$ such that $[q_{\min},q_{\max}] \supseteq Q^m$. So we have the stated properties of the adjusted value function for $q_c \neq \hat{q}$ as well.

Suppose therefore that $q_c = \hat{q}$, implying that the central ray $\mathcal{R}_c$ is vertical. This simplifies the following analysis since it rules out intersections between the graph of v^* and $\mathcal{R}_c$, so we do not have to worry about the 'break' in (19) along $\mathcal{R}_c$.

Below, we will make repeated use of the following observation:

$$(v^*)'(\pi) - 2\Lambda\,(\pi - \tilde{\pi})G(\pi, v^*(\pi))/\tau(\pi) \text{ converges to a finite limit as } \pi \to \hat{\pi}. \qquad \text{(F.1)}$$

In fact, the definition of the adjusted value function in (18) and the ODE (14) together with its undiscounted variant imply that this expression equals $(1+\Lambda/r)(u^*)'(\pi)$ if $r>0$, and $\Lambda(u^*)'(\pi)$ otherwise. So (F.1) follows from continuity of $(u^*)'$.

We can now turn to the proof of convexity of v^*. For the sake of concreteness, we assume that $\hat{\pi}<\tilde{\pi}$. Again, this is without loss of generality, since we could always relabel the demand curves.

We consider the subinterval left of $\hat{\pi}$ first. Suppose that $(v^*)''\leq 0$ on $]0,\hat{\pi}[$. Fix any π in this interval and let w be the function w_k that satisfies $w(\pi)=v^*(\pi)$. Then, $w'(\pi)\leq(v^*)'(\pi)$ by the above lemma. In fact, the equality $w'(\pi)=(v^*)'(\pi)$ is precluded since it would imply convexity of v^* immediately to the right of π (v^* would cross a nearby function w_k from above). So $w'(\pi)<(v^*)'(\pi)$. Since $w(\hat{\pi})>\hat{m}=v^*(\hat{\pi})$, there must be a π' in $]\pi,\hat{\pi}[$ such that $w(\pi')=v^*(\pi')$ and $w'(\pi')>(v^*)'(\pi')$ (equality is again precluded). But this means $(v^*)''(\pi')>0$, a contradiction. Thus, we must have $\inf\{\pi\in]0,\hat{\pi}[\colon (v^*)''(\pi)>0\}<\hat{\pi}$.

Now suppose that this infimum is positive, so there is a belief $\pi^\dagger$ with $0<\pi^\dagger<\hat{\pi}$ and $(v^*)''(\pi^\dagger)\leq 0$. Let $k^\dagger>0$ be such that the function $w^\dagger=w_{k^\dagger}$ coincides with v^* at $\pi^\dagger$. By the above lemma, $(v^*)'(\pi^\dagger)\geq(w^\dagger)'(\pi^\dagger)$. Now, the strict convexity of the functions w_k implies that $(v^*)'(\pi)>w_k'(\pi)$ for $\pi<\pi^\dagger$ sufficiently close to $\pi^\dagger$ and $k<k^\dagger$ sufficiently close to $k^\dagger$. So $(v^*)''$ is strictly negative immediately to the left of $\pi^\dagger$. On the other hand, v^* cannot be strictly concave on the whole of $]0,\pi^\dagger[$ since this would imply $(v^*)'>w'$ on $]0,\pi^\dagger[$ and hence $v^*(0)<w^\dagger(0)=m(0)$. Therefore, there must be a positive $\pi<\pi^\dagger$ such that $(v^*)''(\pi)=0$ again; let $\pi^\ddagger$ be the biggest such π, and $w^\ddagger$ that function w_k which is tangent to v^* at $\pi^\ddagger$. On $[\pi^\ddagger,\pi^\dagger]$, $w^\ddagger$ and $w^\dagger$ are strictly convex, while v^* is strictly concave. This implies $w^\ddagger(\pi^\ddagger)<w^\dagger(\pi^\ddagger)$ and $w^\ddagger(\pi^\dagger)>w^\dagger(\pi^\dagger)$, so $w^\ddagger$ and $w^\dagger$ must intersect somewhere on $]\pi^\ddagger,\pi^\dagger[$ – a contradiction. This proves that $(v^*)''>0$ on $]0,\hat{\pi}[$.

Using the same argument as in the previous paragraph, we also see that $(v^*)''>0$ on $]\tilde{\pi},1[$. Now let $\overline{\pi}=\inf\{\pi>\hat{\pi}:(v^*)''(\pi)>0\}$. We know that $(v^*)''(\tilde{\pi})>0$, so $\hat{\pi}\leq\overline{\pi}<\tilde{\pi}$. Suppose $\overline{\pi}>\hat{\pi}$. Arguing once more as in the previous paragraph, we can show that $(v^*)''<0$ immediately to the left of $\overline{\pi}$. Moreover, $(v^*)''$ must be negative on the whole of $]\hat{\pi},\overline{\pi}[$ since the existence of a π in this interval with $(v^*)''(\pi)\geq 0$ would again lead to a contradiction. Thus, the one-sided derivatives of v^* at $\hat{\pi}$ are well defined, and we must have $(D_-v^*)(\hat{\pi})\leq 0$ and $(D_+v^*)(\hat{\pi})>0$ since $v^*\geq m$, $m'(\hat{\pi})=0$ and v^* is strictly concave immediately to the right of $\hat{\pi}$. In view of (F.1), we conclude that $G(\pi,v^*(\pi))$ has one-sided limits at $\hat{\pi}$ with $\lim_{\pi\to\hat{\pi}-}G(\pi,v^*(\pi))>\lim_{\pi\to\hat{\pi}+}G(\pi,v^*(\pi))$. The explicit representation for G in (16)–(17) shows that these limits lie in the interval $[0,\beta(\hat{\pi})]$. However, $(D_+v^*)(\hat{\pi})>0$ implies $\lim_{\pi\to\hat{\pi}+}G(\pi,v^*(\pi))=\beta(\hat{\pi})$ by L'Hôpital's rule, and hence $\lim_{\pi\to\hat{\pi}-}G(\pi,v^*(\pi))>\beta(\hat{\pi})$ – a contradiction. This proves that $\overline{\pi}=\hat{\pi}$ and $(v^*)''>0$ on the whole of $]0,1[\,-\{\hat{\pi}\}$.

In particular, the one-sided derivatives of v^* at $\hat{\pi}$ exist and satisfy $(D_-v^*)(\hat{\pi}) \leq 0 \leq (D_+v^*)(\hat{\pi})$. Repeating the argument given in the previous paragraph, we would again get a contradiction if $(D_+v^*)(\hat{\pi}) > 0$. So $(D_+v^*)(\hat{\pi}) = 0$.

Observation (F.1) now implies the existence of one-sided limits $\lim_{\pi\to\hat{\pi}-} G(\pi, v^*(\pi)) \geq \lim_{\pi\to\hat{\pi}+} G(\pi, v^*(\pi))$ in the interval $[0, \beta(\hat{\pi})]$, the inequality being strict iff $(D_-v^*)(\hat{\pi}) < 0$. Having established convexity of v^*, we also know that its graph lies entirely in the region associated with interior quantities. So the relevant expression for the function G is $G(\pi, v) = \beta(\pi)[v - m(\pi)]/[v - \hat{m}]$. We can now prove the rest of the proposition.

If $(D_-v^*)(\hat{\pi}) < 0$, then $\lim_{\pi\to\hat{\pi}+} G(\pi, v^*(\pi)) < \lim_{\pi\to\hat{\pi}-} G(\pi, v^*(\pi)) = \beta(\hat{\pi})$ where the equality follows again by L'Hôpital's rule. Since $[v-\hat{m}]/[m(\pi)-\hat{m}] = \beta(\pi)/[\beta(\pi)-G(\pi, v)]$, this proves that $[v^*(\pi) - \hat{m}]/[m(\pi) - \hat{m}]$ tends to a finite limit as $\pi \to \hat{\pi}+$.

If $(D_-v^*)(\hat{\pi}) = 0$, on the other hand, then $G(\pi, v^*(\pi))$ approaches the same limit from both sides of $\hat{\pi}$. If this limit is strictly smaller than $\beta(\hat{\pi})$, then the quotient $[v^*(\pi) - \hat{m}]/[m(\pi) - \hat{m}]$ has a finite limit, as in the previous paragraph. Suppose therefore that $\lim_{\pi\to\hat{\pi}} G(\pi, v^*(\pi)) = \beta(\hat{\pi})$. Then the function $h(\pi) = G(\pi, v^*(\pi))/\beta(\pi) = [v^*(\pi) - m(\pi)]/[v^*(\pi) - \hat{m}]$, which is strictly smaller than 1 for $\pi \neq \hat{\pi}$, tends to 1 as $\pi \to \hat{\pi}$. For every $\pi \neq \hat{\pi}$, we can find a ξ between π and $\hat{\pi}$ such that $m(\pi)-\hat{m} = \frac{1}{2}m''(\xi)(\pi-\hat{\pi})^2$, hence $1 - h(\pi) = [m(\pi) - \hat{m}]/[v^*(\pi) - \hat{m}] = \frac{1}{2}m''(\xi)(\pi - \hat{\pi})^2/[v^*(\pi) - \hat{m}]$. Since $(v^*)'(\hat{\pi}) = 0$, the ratio $[1 - h(\pi)]/(\hat{\pi} - \pi)$ is unbounded above as π approaches $\hat{\pi}$ from the left and, by the mean value theorem, so is $h'(\pi)$. As $G(\pi, v^*(\pi)) = \beta(\pi)h(\pi)$, we conclude that $(d/d\pi)G(\pi, v^*(\pi))$ is also unbounded above. But given that $\hat{\pi} < \tilde{\pi}$, (19) now implies that $(v^*)''$ is unbounded below – a contradiction. ∎

We still have to show that the functions w_k are themselves convex.

Lemma F.2 *For $k > 0$, the function w_k is strictly convex at all $\pi \in]0, 1[- \{\tilde{\pi}\}$ such that $(\pi, w_k(\pi)) \notin \mathcal{R}_c$.*

PROOF: We fix a $k > 0$ and simply write w for the corresponding function w_k.

We first consider π such that $(\pi, w(\pi))$ lies in one of the regions associated with an interior quantity. It is straightforward to show that in these regions, w satisfies the first-order ODE

$$w'(\pi) = \frac{w(\pi) - \hat{m}}{m(\pi) - \hat{m}} \left[m'(\pi) - K(\pi)\left(w(\pi) - m(\pi)\right)\right]$$

with

$$K(\pi) = \frac{2 + r/\Lambda}{\pi - \tilde{\pi}} + \frac{\Delta\beta}{\beta(\pi)} - \frac{2(1 - 2\pi)}{\pi(1 - \pi)}\,.$$

Differentiating both sides and using the ODE to replace $w'(\pi)$, we find that $w''(\pi)$ is a quadratic in $w(\pi) - m(\pi)$ multiplied by a positive factor:

$$w'' = \Big[a\,(w - m)^2 + b\,(w - m) + c\Big]\,d$$

with

$$\begin{aligned} a &= 2K^2, \\ b &= (K^2 - K')(m - \hat{m}) - 2Km', \\ c &= (m - \hat{m})m'', \\ d &= (w - \hat{m})/(m - \hat{m})^2, \end{aligned}$$

where we have suppressed the dependence of the functions K, m and w on π. Clearly $a > 0$ and, since $m > \hat{m}$ and m is convex, we also have $c > 0$. Thus, if $b \geq 0$, then $w'' > 0$ and we are done.

Suppose therefore that $b < 0$. We have to show that the above quadratic in $w-m$ has no real roots. This is the same as showing $b^2 - 4ac < 0$, or equivalently $(b+2\sqrt{ac})(b-2\sqrt{ac}) < 0$. Since we are dealing with the case $b < 0$, the second factor is negative, so all we have to show is that $b + 2\sqrt{ac} > 0$, i.e.,

$$(K^2 - K')(m - \hat{m}) - 2Km' + 2\sqrt{2K^2(m - \hat{m})m''} > 0$$

when $(K^2 - K')(m - \hat{m}) - 2Km' < 0$.

First, we can show that $K^2 - K' > 0$. Indeed, $K^2 - K' = [(r/\Lambda)^2 + (r/\Lambda)f_1 + 2f_2]/(\pi - \tilde{\pi})^2$ with

$$\begin{aligned} f_1(\pi) &= \{\tilde{\pi}(1-\pi)\left[(1-\pi)\left(\beta_0 + \beta(\pi)\right) + (1-\pi)\beta_0 + \beta(\pi)\right] \\ &\quad + (1-\tilde{\pi})\pi\left[\pi\left(\beta_1 + \beta(\pi)\right) + \pi\beta_1 + \beta(\pi)\right]\}/\{\pi(1-\pi)\beta(\pi)\} \end{aligned}$$

and

$$f_2(\pi) = \frac{\left(\beta_0\tilde{\pi}(1-\pi)^2 + \beta_1(1-\tilde{\pi})\pi^2\right)^2 + 2\tilde{\pi}(1-\tilde{\pi})\pi(1-\pi)\beta(\pi)^2}{\left(\pi(1-\pi)\beta(\pi)\right)^2}$$

which are positive by inspection. This leads to the further simplification that the only possibility we need consider is when $Km' > 0$.

It will be more convenient to rewrite $m(\pi)$ and its derivatives in terms of $q^m(\pi)$ and $\hat{q}$ as follows:

$$m - \hat{m} = \beta\left(q^m - \hat{q}\right)^2, \qquad m' = -\Delta\beta\, q^m\left(q^m - \hat{q}\right), \qquad m'' = \frac{\Delta\beta^2}{2\beta}\left(2q^m - \hat{q}\right)^2$$

where again we have suppressed the dependence of the functions β, m and q^m on π. Having made these substitutions, the expression which we wish to show is positive becomes

$$\beta\left(K^2 - K'\right)\left(q^m - \hat{q}\right)^2 + 2\,\Delta\beta\,\left\{Kq^m\left(q^m - \hat{q}\right) + |K|\,|q^m - \hat{q}|\,|2q^m - \hat{q}|\right\},$$

and we need consider only the possibility that $K\,(q^m - \hat{q}) < 0$.

(1) $K < 0$, $q^m - \hat{q} > 0$. The expression in braces becomes

$$-|K|\, q^m \,(q^m - \hat{q}) + |K|\,(q^m - \hat{q})\,(2q^m - \hat{q}) = |K|\,(q^m - \hat{q})^2\,,$$

so we are done.

(2) $K > 0$, $q^m - \hat{q} < 0$. This time the expression in braces becomes

$$-Kq^m\,|q^m - \hat{q}| + K\,|q^m - \hat{q}|\,|2q^m - \hat{q}| = K\,|q^m - \hat{q}|\,(|2q^m - \hat{q}| - q^m)\,,$$

and we have two subcases.

(2i) $2q^m \geq \hat{q}$. The above expression becomes

$$K\,|q^m - \hat{q}|\,(2q^m - \hat{q} - q^m) = -K\,(q^m - \hat{q})^2$$

and the whole expression which we wish to show is positive becomes

$$\beta\,(K^2 - K')\,(q^m - \hat{q})^2 - 2\,\Delta\beta\,K\,(q^m - \hat{q})^2 = \Big[\beta\,(K^2 - K') - 2\,\Delta\beta\,K\Big]\,(q^m - \hat{q})^2\,.$$

(2ii) $2q^m < \hat{q}$. Now the expression for case (2) becomes

$$K\,|q^m - \hat{q}|\,(-2q^m + \hat{q} - q^m) = K\,(\hat{q} - q^m)\,(\hat{q} - 3q^m)$$

and the whole expression which we wish to show is positive becomes

$$\left[\beta\,(K^2 - K') + 2\,\Delta\beta\,K\frac{\hat{q} - 3q^m}{\hat{q} - q^m}\right]\,(q^m - \hat{q})^2\,.$$

But $q^m - \hat{q} < 0$ and $2q^m < \hat{q}$ imply that $(\hat{q} - 3q^m)/(\hat{q} - q^m) > -1$, and so the above expression is greater than

$$\Big[\beta\,(K^2 - K') - 2\,\Delta\beta\,K\Big]\,(q^m - \hat{q})^2$$

which is just the expression that we found in case (2i). Therefore, case (2) comes down to showing that the term in square brackets is positive. In fact, it can be written as

$$\begin{aligned}
&\frac{2\beta(\pi)\left[\tilde{\pi}^2(1-\pi)^4 + (1-\tilde{\pi})^2\pi^4 + 2\tilde{\pi}(1-\tilde{\pi})\pi(1-\pi)\,(1+\pi(1-\pi))\right]}{\pi^2(1-\pi)^2(\pi-\tilde{\pi})^2} \\
+\quad &\frac{\beta(\pi)(r/\Lambda)\left[\tilde{\pi}(1-\pi)\,(1+3(1-\pi)) + (1-\tilde{\pi})\pi\,(1+3\pi)\right]}{\pi(1-\pi)(\pi-\tilde{\pi})^2} \\
+\quad &\frac{\beta(\pi)(r/\Lambda)^2}{(\pi-\tilde{\pi})^2}
\end{aligned}$$

which is positive by inspection.

Thus $w''(\pi) > 0$ in the regions associated with interior solutions.

We still have to consider π such that $(\pi, w(\pi))$ lies in a region associated with exactly one of the extreme quantities. This quantity, which can be either $q_{\max}$ or $q_{\min}$, will be denoted by $q^\dagger$. It is straightforward to show that in such a region, w satisfies

$$w'(\pi) = q^\dagger \left(\Delta\alpha - \Delta\beta\, q^\dagger\right) - L(\pi) \left(w(\pi) - R(\pi, q^\dagger)\right)$$

with

$$L(\pi) = \frac{2 + r/\Lambda}{\pi - \tilde{\pi}} - \frac{2(1-2\pi)}{\pi(1-\pi)}\,,$$

and that

$$w''(\pi) = \left[L^2(\pi) - L'(\pi)\right] \left(w(\pi) - R(\pi, q^\dagger)\right).$$

As $w(\pi) > R(\pi, q^\dagger)$ by construction, we only have to show that $L^2 - L' > 0$. This follows from the representation $L^2 - L' = [(r/\Lambda)^2 + (r/\Lambda)g_1 + 2g_2]/(\pi - \tilde{\pi})^2$ where

$$g_1(\pi) = \frac{\tilde{\pi}(1-\pi)\left[3(1-\pi)+1\right] + (1-\tilde{\pi})\pi\left[3\pi+1\right]}{\pi(1-\pi)}$$

and

$$g_2(\pi) = \frac{\left(\tilde{\pi}(1-\pi)^2 + (1-\tilde{\pi})\pi^2\right)^2 + 2\tilde{\pi}(1-\tilde{\pi})\pi(1-\pi)}{\pi^2(1-\pi)^2}$$

are clearly positive; this representation is obtained by setting $\Delta\beta = 0$ in the above expression for $K^2 - K'$. ∎

G Two-Point Boundary Value Problems

Consider a second-order differential equation of the form

$$\pi^2(1-\pi)^2\, v'' = F[\pi, v, v'] \tag{G.1}$$

on some open interval $I =]\pi_\ell, \pi_r[\subseteq]0,1[$. We are interested in finding a solution to this ODE which assumes prespecified values at the two boundary points of the interval.

The existence theorem presented below requires the concept of a sub- or supersolution to this ODE. Let v be a real-valued continuous function on $\overline{I} = [\pi_\ell, \pi_r]$ with a continuous first derivative on I. Define functions $\underline{D}v', \overline{D}v' : I \to \mathbb{R} \cup \{\pm\infty\}$ by

$$(\underline{D}v')(\pi) = \liminf_{h\to 0} \frac{v'(\pi+h) - v'(\pi-h)}{2h}\,, \quad (\overline{D}v')(\pi) = \limsup_{h\to 0} \frac{v'(\pi+h) - v'(\pi-h)}{2h}\,.$$

(Note that for twice differentiable v, the functions $\underline{D}v'$ and $\overline{D}v'$ coincide with v''.) The function v is called a *subsolution* of the ODE (G.1) if $\pi^2(1-\pi)^2\, \underline{D}v' \geq F[\pi, v, v']$

on I. Similarly, v is called a *supersolution* if $\pi^2(1-\pi)^2\,\overline{D}v' \leq F[\pi, v, v']$ on I. We speak of a *strict* subsolution or supersolution if the respective inequality is strict on I.

Fix functions $\underline{v}, \overline{v} : \overline{I} \to I\!R$ satisfying $\underline{v} \leq \overline{v}$ on $\overline{I}$. Given any subinterval $J \subseteq \overline{I}$, we say that the function F on the right-hand side of (G.1) is *regular on J with respect to $\underline{v}$ and $\overline{v}$* if it is continuous on

$$\mathcal{D}_J = \{(\pi, v_0, v_1) \in J \times I\!R \times I\!R : \underline{v}(\pi) \leq v_0 \leq \overline{v}(\pi)\}$$

and there is a constant C_J depending only on J such that $|F[\pi, v_0, v_1]| \leq C_J\,(1 + |v_1|)$ on $\mathcal{D}_J$.

Proposition G.1 *Let $0 < \pi_\ell < \pi_r < 1$. Assume that $\underline{v} : \overline{I} \to I\!R$ is a subsolution of* (G.1)*, $\overline{v} : \overline{I} \to I\!R$ a supersolution, and $\underline{v} \leq \overline{v}$. If F is regular with respect to $\underline{v}$ and $\overline{v}$ on $\overline{I} = [\pi_\ell, \pi_r]$, then for any $v_\ell \in [\underline{v}(\pi_\ell), \overline{v}(\pi_\ell)]$ and $v_r \in [\underline{v}(\pi_r), \overline{v}(\pi_r)]$, there is a continuous function $v : \overline{I} \to I\!R$ which solves* (G.1) *on I with $\underline{v} \leq v \leq \overline{v}$ and satisfies the boundary conditions $v(\pi_\ell) = v_\ell$ and $v(\pi_r) = v_r$. Moreover, if $\underline{v}$ is a strict subsolution, then $v > \underline{v}$ on I, and if $\overline{v}$ is a strict supersolution, then $v < \overline{v}$ on I.*

PROOF: The existence of such a solution v follows directly from Bernfeld and Lakshmikantham (1974, Theorem 1.5.1). Now assume that $\underline{v}$ is a strict subsolution and that there is a belief $\breve{\pi} \in I$ such that $v(\breve{\pi}) = \underline{v}(\breve{\pi})$. Then the function $v - \underline{v}$ has a local minimum at $\breve{\pi}$, so $v'(\breve{\pi}) = \underline{v}'(\breve{\pi})$ and $v''(\breve{\pi}) \geq (\underline{D}\,\underline{v}')(\breve{\pi})$. Yet $\breve{\pi}^2(1-\breve{\pi})^2\,v''(\breve{\pi}) = F[\breve{\pi}, v(\breve{\pi}), v'(\breve{\pi})] = F[\breve{\pi}, \underline{v}(\breve{\pi}), \underline{v}'(\breve{\pi})] < \breve{\pi}^2(1-\breve{\pi})^2\,(\underline{D}\,\underline{v}')(\breve{\pi})$ – a contradiction. The case of a strict supersolution $\overline{v}$ is dealt with in the same way. ∎

We will need the following corollary of this result.

Corollary G.1 *Given $\pi_\ell < \pi_c < \pi_r$ in $]0, 1[$, consider the ODEs*

$$\pi^2(1-\pi)^2\,v'' = F_\ell[\pi, v, v'] \tag{G.2}$$

on $]\pi_\ell, \pi_c[$ and

$$\pi^2(1-\pi)^2\,v'' = F_r[\pi, v, v'] \tag{G.3}$$

on $]\pi_c, \pi_r[$. Let $\underline{v}_\ell : [\pi_\ell, \pi_c] \to I\!R$ be a subsolution of (G.2)*, $\overline{v}_\ell : [\pi_\ell, \pi_c] \to I\!R$ a supersolution of* (G.2)*, $\underline{v}_r : [\pi_c, \pi_r] \to I\!R$ a subsolution of* (G.3) *and $\overline{v}_r : [\pi_c, \pi_r] \to I\!R$ a supersolution of* (G.3) *such that $\underline{v}_\ell \leq \overline{v}_\ell$, $\underline{v}_r \leq \overline{v}_r$, $\underline{v}_\ell(\pi_c) = \underline{v}_r(\pi_c) < \overline{v}_\ell(\pi_c) = \overline{v}_r(\pi_c)$, $\underline{v}'_\ell(\pi_c-) \leq \underline{v}'_r(\pi_c+)$ and $\overline{v}'_\ell(\pi_c-) \geq \overline{v}'_r(\pi_c+)$. Assume that F_ℓ is regular with*

respect to $\underline{v}_\ell$ *and* $\overline{v}_\ell$ *on each closed interval contained in* $]\pi_\ell, \pi_c]$*, and* F_r *is regular with respect to* $\underline{v}_r$ *and* $\overline{v}_r$ *on each closed interval contained in* $[\pi_c, \pi_r[$*. Then there is a differentiable function* $v :]\pi_\ell, \pi_r[\to \mathbb{R}$ *which solves* (G.2) *on* $]\pi_\ell, \pi_c[$ *and* (G.3) *on* $]\pi_c, \pi_r[$ *such that* $\underline{v}_\ell \leq v \leq \overline{v}_\ell$ *on* $]\pi_\ell, \pi_c]$ *and* $\underline{v}_r \leq v \leq \overline{v}_r$ *on* $[\pi_c, \pi_r[$*.*

PROOF: Piecing together $\underline{v}_\ell$, $\underline{v}_r$ and $\overline{v}_\ell$, $\overline{v}_r$ in the obvious way, we get continuous functions $\underline{v}, \overline{v} : [\pi_\ell, \pi_r] \to \mathbb{R}$. Let $\epsilon > 0$ be such that $\pi_\ell + \epsilon < \pi_c < \pi_r - \epsilon$. We shall construct numbers $\underline{a}_n, \overline{a}_n \in [\underline{v}(\pi_c), \overline{v}(\pi_c)]$ and functions $\underline{v}_n, \overline{v}_n$ with the following properties for all $n = 1, 2, \ldots$:

(i) $\underline{a}_n < \overline{a}_n$;

(ii) $\underline{a}_{n+1} \geq \underline{a}_n$ and $\overline{a}_{n+1} \leq \overline{a}_n$;

(iii) $\underline{v}_n, \overline{v}_n : [\pi_\ell + \epsilon, \pi_r - \epsilon] \to \mathbb{R}$ are continuous and solve (G.2) on $]\pi_\ell + \epsilon, \pi_c[$ and (G.3) on $]\pi_c, \pi_r - \epsilon[$ subject to $\underline{v}_n(\pi_\ell + \epsilon) = \overline{v}_n(\pi_\ell + \epsilon) = \overline{v}(\pi_\ell + \epsilon)$, $\underline{v}_n(\pi_r - \epsilon) = \overline{v}_n(\pi_r - \epsilon) = \overline{v}(\pi_r - \epsilon)$, $\underline{v}_n(\pi_c) = \underline{a}_n$, $\overline{v}_n(\pi_c) = \overline{a}_n$;

(iv) $\underline{v}'_n(\pi_c-) \leq \underline{v}'_n(\pi_c+)$ and $\overline{v}'_n(\pi_c-) \geq \overline{v}'_n(\pi_c+)$;

(v) $\underline{v} \leq \underline{v}_n < \overline{v}_n \leq \overline{v}$ on $]\pi_\ell + \epsilon, \pi_r - \epsilon[$;

(vi) $\underline{v}_{n+1} \geq \underline{v}_n$ and $\overline{v}_{n+1} \leq \overline{v}_n$.

For $n = 1$, we set $\underline{a}_1 = \underline{v}(\pi_c)$ and $\overline{a}_1 = \overline{v}(\pi_c)$, so (i) holds. Using Proposition G.1 separately to the left and right of π_c, we find a function $\underline{v}_1$ satisfying (iii) and $\underline{v} \leq \underline{v}_1 \leq \overline{v}$, and a function $\overline{v}_1$ satisfying (iii) and $\underline{v}_1 \leq \overline{v}_1 \leq \overline{v}$. Property (iv) is then obvious, and a simple argument similar to the one given at the end of the previous proof shows (v). Suppose we have constructed $\underline{a}_n, \overline{a}_n, \underline{v}_n$ and $\overline{v}_n$ with (i) and (iii)–(v). If $\underline{v}'_n(\pi_c-) = \underline{v}'_n(\pi_c+)$ or $\overline{v}'_n(\pi_c-) = \overline{v}'_n(\pi_c+)$, we simply set $\underline{a}_{n+1} = \underline{a}_n$, $\overline{a}_{n+1} = \overline{a}_n$, $\underline{v}_{n+1} = \underline{v}_n$ and $\overline{v}_{n+1} = \overline{v}_n$. Otherwise, we consider $a = (\underline{a}_n + \overline{a}_n)/2$ and a continuous function $v : [\pi_\ell + \epsilon, \pi_r - \epsilon] \to \mathbb{R}$ which satisfies $\underline{v}_n \leq v \leq \overline{v}_n$ and solves (G.2) on $]\pi_\ell + \epsilon, \pi_c[$ and (G.3) on $]\pi_c, \pi_r - \epsilon[$ subject to $v(\pi_\ell + \epsilon) = \overline{v}(\pi_\ell + \epsilon)$, $v(\pi_r - \epsilon) = \overline{v}(\pi_r - \epsilon)$, $v(\pi_c) = a$. Such a function exists by Proposition G.1, and it is again straightforward to see that $\underline{v}_n < v < \overline{v}_n$ on $]\pi_\ell + \epsilon, \pi_r - \epsilon[$. If $v'(\pi_c-) \leq v'(\pi_c+)$, we set $\underline{a}_{n+1} = a$, $\overline{a}_{n+1} = \overline{a}_n$, $\underline{v}_{n+1} = v$ and $\overline{v}_{n+1} = \overline{v}_n$; if $v'(\pi_c-) > v'(\pi_c+)$, we set $\underline{a}_{n+1} = \underline{a}_n$, $\overline{a}_{n+1} = a$, $\underline{v}_{n+1} = \underline{v}_n$ and $\overline{v}_{n+1} = v$. This procedure clearly implies (i)–(vi).

If none of the functions $\underline{v}_n$ or $\overline{v}_n$ is differentiable at π_c, then the sequences $(\underline{a}_n)$ and $(\overline{a}_n)$ converge to a common limit a_∞, and by Bernfeld and Lakshmikantham (1974, Corollary 1.5.1), the sequences $(\underline{v}_n)$ and $(\overline{v}_n)$ have subsequences converging uniformly to functions $\underline{v}_\infty \leq \overline{v}_\infty$ which solve (G.2) and (G.3) on the respective open intervals, with the corresponding subsequences of $(\underline{v}'_n)$ and $(\overline{v}'_n)$ converging to $\underline{v}'_\infty$ and $\overline{v}'_\infty$, respectively. As $\underline{v}_\infty(\pi_c) = \overline{v}_\infty(\pi_c) = a_\infty$, we have $\underline{v}'_\infty(\pi_c-) \geq \overline{v}'_\infty(\pi_c-)$ and $\underline{v}'_\infty(\pi_c+) \leq \overline{v}'_\infty(\pi_c+)$. On the other hand, $\underline{v}'_\infty(\pi_c-) \leq \underline{v}'_\infty(\pi_c+)$ and $\overline{v}'_\infty(\pi_c-) \geq \overline{v}'_\infty(\pi_c+)$, hence $\underline{v}'_\infty(\pi_c-) \geq \overline{v}'_\infty(\pi_c-) \geq \overline{v}'_\infty(\pi_c+) \geq \underline{v}'_\infty(\pi_c+) \geq \underline{v}'_\infty(\pi_c-)$, implying equality throughout.

For any small $\epsilon > 0$, we can therefore always find a function v_ϵ on $[\pi_\ell + \epsilon, \pi_r - \epsilon]$ which solves (G.2) on $]\pi_\ell + \epsilon, \pi_c[$ and (G.3) on $]\pi_c, \pi_r - \epsilon[$, is differentiable at π_c, and satisfies $v_\epsilon(\pi_\ell + \epsilon) = \overline{v}(\pi_\ell + \epsilon)$, $v_\epsilon(\pi_r - \epsilon) = \overline{v}(\pi_r - \epsilon)$, and $\underline{v} \leq v_\epsilon \leq \overline{v}$ everywhere else.

Finally, consider a sequence $v_k : [\pi_\ell + \epsilon_k, \pi_r - \epsilon_k] \to I\!R$ of such functions for small positive numbers $(\epsilon_k)_{k=1,2,\ldots}$ converging monotonically to 0. By Bernfeld and Lakshmikantham (1974, Theorem 1.4.1), there is an $N_k > 0$ such that $|v'| \leq N_k$ on $[\pi_\ell + \epsilon_k, \pi_r - \epsilon_k]$ for *any* solution lying between $\underline{v}$ and $\overline{v}$ on this interval. Thus for any fixed integer $K \geq 1$ and all $k \geq K$, v_k is a solution satisfying $\underline{v} \leq v_k \leq \overline{v}$ and $|v_k'| \leq N_K$ on $[\pi_\ell + \epsilon_K, \pi_r - \epsilon_K]$, so the sequences $(v_k)_{k \geq K}$ and $(v_k')_{k \geq K}$ are both uniformly bounded and equicontinuous on that interval. Employing the standard diagonalisation argument, we obtain a subsequence which converges uniformly on all compact subintervals of $]\pi_\ell, \pi_r[$ to a function v with the desired properties. ∎

Our analysis of the Bellman equation lead us to the following second-order differential equation for the adjusted value function v^*:

$$\tau(\pi)\frac{v''(\pi)}{2} = r\,G(\pi, v(\pi)) + \Lambda\left\{f(\pi)\,G(\pi, v(\pi)) + (\pi - \tilde{\pi})\frac{d}{d\pi}G(\pi, v(\pi))\right\} \quad \text{(G.4)}$$

with the positive function

$$f(\pi) = 2\left(\frac{\tilde{\pi}(1-\pi)}{\pi} + \frac{(1-\tilde{\pi})\pi}{1-\pi}\right)$$

and G defined by equation (13). We saw that G is continuously differentiable in the area $\mathcal{A}$ with the exception of the central ray in case $\hat{q}$ lies in the interior of Q^m; if this is the case, we consider the ODE separately to the left and to the right of the central ray. Throughout, we will assume that at least one of the parameters r and Λ is strictly positive.

Lemma G.1 *The myopic pay-off function m is a strict subsolution of* (G.4) *on* $]0,1[$ *if $\hat{q}$ is not in the interior of Q^m, and on* $]0,1[- \{\hat{\pi}\}$ *otherwise.*

PROOF: $m'' > 0$, and we have $G(\pi, m(\pi)) = 0$ on the stated sets of beliefs. ∎

Recall that the full-information pay-off function is defined by $\overline{m}(\pi) = (1 - \pi)\,m(0) + \pi\,m(1)$. Its graph is the straight line joining $(0, m(0))$ and $(1, m(1))$.

Lemma G.2 *The full information pay-off function $\overline{m}$ is a strict supersolution of* (G.4).

PROOF: $\overline{m}'' = 0$, so we have to show that the right-hand side of (G.4) with $v(\pi)$ replaced by $\overline{m}(\pi) = (1 - \pi)\,m(0) + \pi\,m(1)$ is positive.

Suppose first that $(\pi, \overline{m}(\pi))$ lies to the left of $\mathcal{R}_\ell$ or to the right of $\mathcal{R}_r$. Then that right-hand side becomes

$$\begin{aligned} r\,\beta(\pi)\overline{H}(\pi) &+ \lambda_0\,(1-\pi)\left[\frac{\beta_0+\beta(\pi)}{\pi}\,\overline{H}(\pi) - \beta(\pi)\,\overline{H}'(\pi)\right] \\ &+ \lambda_1\,\pi\left[\frac{\beta_1+\beta(\pi)}{1-\pi}\,\overline{H}(\pi) + \beta(\pi)\,\overline{H}'(\pi)\right] \end{aligned}$$

with $\overline{H}(\pi) = [\overline{m}(\pi) - m(\pi)]/[\overline{m}(\pi) - \hat{m}]$. The first term is clearly positive. The expressions in square brackets associated with λ_0 and λ_1 simplify to $h_0(\pi)/(\overline{m}(\pi)-\hat{m})^2$ and $h_1(\pi)/(\overline{m}(\pi)-\hat{m})^2$ respectively, where h_0 and h_1 are quadratics in π:

$$\begin{aligned} h_0(\pi) &= K\,\left[m(1) - \hat{m} + [m(0)-m(1)]\,(1-\pi)^2\right], \\ h_1(\pi) &= K\,\left[m(0) - \hat{m} + [m(1)-m(0)]\,\pi^2\right] \end{aligned}$$

with $K = \beta_0\,\beta_1\,[q^m(0) - q^m(1)]^2$. Thus, $h_0(0) = h_1(0) = K\,[m(0)-\hat{m}]$ and $h_0(1) = h_1(1) = K\,[m(1)-\hat{m}]$, so h_0 and h_1 are both non-negative at each end of the unit interval. As the two quadratics are strictly monotonic on $[0,1]$, they are both non-negative over the entire unit interval.

Next consider π such that $(\pi, \overline{m}(\pi))$ lies between the rays $\mathcal{R}_\ell$ and $\mathcal{R}_r$. In such a region, the right-hand side for $\overline{m}$ can be written as

$$r\,\overline{G}(\pi) + \lambda_0\,(1-\pi)\left[\frac{2}{\pi}\,\overline{G}(\pi) - \overline{G}'(\pi)\right] + \lambda_1\,\pi\left[\frac{2}{1-\pi}\,\overline{G}(\pi) + \overline{G}'(\pi)\right]$$

where $\overline{G}(\pi) = (\overline{m}(\pi) - m(\pi) + \beta(\pi)[q^\dagger - q^m(\pi)]^2)/[q^\dagger - \hat{q}]^2$ and the quantity $q^\dagger$ is either $q_{\max}$ or $q_{\min}$. Again, the first term is clearly positive. The expressions in square brackets associated with λ_0 and λ_1 simplify to $\ell_0(\pi)/(\pi\,[q^\dagger - \hat{q}]^2)$ and $\ell_1(\pi)/((1-\pi)[q^\dagger - \hat{q}]^2)$ respectively, where ℓ_0 and ℓ_1 are the following linear functions:

$$\begin{aligned} \ell_0(\pi) &= \left(\beta_0\,[q^\dagger - q^m(0)]^2 + \beta_1\,[q^\dagger - q^m(1)]^2\right) \\ &\quad + \left(\beta_0\,[q^\dagger - q^m(0)]^2 - \beta_1\,[q^\dagger - q^m(1)]^2\right)\,(1-\pi)\,, \\ \ell_1(\pi) &= \left(\beta_1\,[q^\dagger - q^m(1)]^2 + \beta_0\,[q^\dagger - q^m(0)]^2\right) \\ &\quad + \left(\beta_1\,[q^\dagger - q^m(1)]^2 - \beta_0\,[q^\dagger - q^m(0)]^2\right)\,\pi\,. \end{aligned}$$

By inspection, these functions are positive on the unit interval. ∎

Define

$$\overline{m}_\ell(\pi) = \frac{\hat{\pi} - \pi}{\hat{\pi}}\,m(0) + \frac{\pi}{\hat{\pi}}\,\hat{m}$$

for $0 \le \pi \le \hat{\pi}$, and

$$\overline{m}_r(\pi) = \frac{1-\pi}{1-\hat{\pi}}\hat{m} + \frac{\pi - \hat{\pi}}{1-\hat{\pi}}\,m(1)$$

for $\hat{\pi} \leq \pi \leq 1$. The graphs of these functions are the rays joining $(\hat{\pi}, \hat{m})$ with $(0, m(0))$ and $(1, m(1))$, respectively.

Lemma G.3 *Let $\hat{q}$ lie in the interior of Q^m. Then the functions $\overline{m}_\ell : [0, \hat{\pi}] \to \mathbb{R}$ and $\overline{m}_r : [\hat{\pi}, 1] \to \mathbb{R}$ are strict supersolutions of* (G.4).

PROOF: The functions $\overline{m}_\ell$ and $\overline{m}_r$ are linear, so $\overline{m}_\ell'' = 0$ and $\overline{m}_r'' = 0$, and their graphs lie entirely in the sub-regions of $\mathcal{A}$ associated with interior solutions. A slightly more complicated variant of the algebra in the first part of the previous proof shows that the right-hand side of (G.4) is positive for these functions. ∎

Lemma G.4 *If $\hat{q}$ does not lie in the interior of Q^m, then the right-hand side of the ODE* (G.4) *is regular with respect to m and $\overline{m}$ on each closed interval contained in $]0, 1[$. Otherwise, the right-hand side of the ODE is regular with respect to m and $\overline{m}_\ell$ on each closed interval contained in $]0, \hat{\pi}[$, and regular with respect to m and $\overline{m}_r$ on each closed interval contained in $]\hat{\pi}, 1[$.*

PROOF: This follows directly from the fact that in the regions associated with interior solutions, the ODE (G.4) is equivalent to the equation $\tau(\pi)v''(\pi)/2 = F[\pi, v(\pi), v'(\pi)]$ where

$$F[\pi, v_0, v_1] = \beta(\pi) \left\{ \left(r + \Lambda \left[f(\pi) + (\pi - \tilde{\pi}) \frac{\Delta\beta}{\beta(\pi)} \right] \right) H(\pi, v_0) + \Lambda \, (\pi - \tilde{\pi}) \, H_1[\pi, v_0, v_1] \right\}$$

with

$$H(\pi, v_0) = \frac{v_0 - m(\pi)}{v_0 - \hat{m}}, \qquad H_1[\pi, v_0, v_1] \; = \; \frac{m(\pi) - \hat{m}}{(v_0 - \hat{m})^2} \; v_1 \; - \; \frac{m'(\pi)}{v_0 - \hat{m}} \; .$$

In particular, v_1 enters linearly. ∎

Proposition G.2 *Suppose that $\hat{q}$ does not lie in the interior of Q^m. Then there is a continuous function $v : [0, 1] \to \mathbb{R}$ which solves* (G.4) *on $]0, 1[$ with $v(0) = m(0)$, $v(1) = m(1)$, and $m < v < \overline{m}$ on $]0, 1[$.*

PROOF: This follows from Lemmas G.1, G.2 and G.4 and Corollary G.1 applied with $\pi_\ell = 0$, $\pi_r = 1$ and $F_\ell = F_r = F$ as given in the proof of Lemma G.4. ∎

Proposition G.3 *Suppose that $\hat{q}$ lies in the interior of Q^m, and fix $\tilde{\pi} \in]0, 1[- \{\hat{\pi}\}$. Then there are positive constants c_1 and c_2 such that for all $r \geq 0$, $\Lambda \geq 0$ and $\sigma > 0$ satisfying $r + c_1\Lambda \geq c_2/\sigma^2$, there exists a continuous function $v : [0, 1] \to \mathbb{R}$ which solves* (G.4) *on $]0, 1[- \{\hat{\pi}\}$ with the following properties: $m < v < \overline{m}_\ell$ on $]0, \hat{\pi}[$, $m < v < \overline{m}_r$ on $]\hat{\pi}, 1[$, and $v - \hat{m} \leq 2\,(m - \hat{m})$ in a neighbourhood of $\hat{\pi}$. In particular, $v(\pi) = m(\pi)$ at $\pi = 0, \; \hat{\pi}, \; 1$ and $v > m$ everywhere else; moreover, v is differentiable with $v'(\hat{\pi}) = 0$.*

PROOF: Consider the function $\breve{m}$ defined by $\breve{m}(\pi) = 2\,m(\pi) - \hat{m}$. Let $\breve{\pi}_\ell$ be the belief where the graph of $\breve{m}$ intersects the graph of $\overline{m}_\ell$, and $\breve{\pi}_r$ the belief where the graph of $\breve{m}$ intersects the graph of $\overline{m}_r$. Define

$$c_1 = \min_{\breve{\pi}_\ell \le \pi \le \breve{\pi}_r} \left[f(\pi) + (\pi - \tilde{\pi})\Delta\beta/\beta(\pi) \right], \qquad c_2 = 2\,\Delta\beta^2 \max_{\breve{\pi}_\ell \le \pi \le \breve{\pi}_r} \frac{\pi^2\,(1-\pi)^2\,m''(\pi)}{\beta(\pi)}\,.$$

While c_2 is clearly positive, the positivity of c_1 follows from the identity

$$f(\pi) + (\pi - \tilde{\pi})\frac{\Delta\beta}{\beta(\pi)} = \frac{\tilde{\pi}(1-\pi)^2\,(\beta_0 + \beta(\pi)) + (1-\tilde{\pi})\pi^2\,(\beta_1 + \beta(\pi))}{\pi(1-\pi)\beta(\pi)}\,.$$

Now let $r + c_1\Lambda \ge c_2/\sigma^2$, implying that $\breve{m}$ is a supersolution of (G.4).

Clearly, the right-hand side of (G.4) is regular with respect to m and $\breve{m}$ on each closed subinterval of $[\breve{\pi}_\ell, \breve{\pi}_r] - \{\hat{\pi}\}$. Since $\breve{m}'(\breve{\pi}_\ell) < \overline{m}'_\ell(\breve{\pi}_\ell)$ and $\breve{m}'(\breve{\pi}_r) > \overline{m}'_r(\breve{\pi}_r)$, Lemma G.3 and Corollary G.1, applied separately to the left and right of $\hat{\pi}$, yield a continuous function $v :]0,1[\to I\!R$ which solves (G.4) on $]0,1[\,-\{\hat{\pi}\}$ with $m \le v \le \overline{m}_\ell$ on $]0,\hat{\pi}]$, $m \le v \le \overline{m}_r$ on $[\hat{\pi},1[$, and $m \le v \le \breve{m}$ on $[\breve{\pi}_\ell, \breve{\pi}_r]$. This function extends continuously to the boundaries of $[0,1]$, and the same argument as in the proof of Proposition G.1 shows that the first and second of these inequalities are strict on $]0,1[\,-\{\hat{\pi}\}$. ∎

Proposition G.4 *Suppose that $\hat{q}$ lies in the interior of Q^m and equals $q_c = (q_{\max} + q_{\min})/2$. Then there are positive constants c_3, c_4 and c_5 such that for all $r \ge 0$, $\Lambda \ge 0$ and $\sigma > 0$ satisfying $c_3 r + c_4\Lambda \le c_5/\sigma^2$, there exists a continuous function $v : [0,1] \to I\!R$ which solves* (G.4) *on $]0,1[\,-\{\hat{\pi}\}$ with the following properties: v is once continuously differentiable and $m < v < \overline{m}$ on $]0,1[$. In particular, $v(\pi) = m(\pi)$ at $\pi = 0,\ 1$ and $v > m$ everywhere else.*

PROOF: Choose a strictly convex function $\underline{m} : [0,1] \to I\!R$ with the following properties: $\underline{m}(0) = m(0)$ and $\underline{m}(1) = m(1)$; $\underline{m} = m$ on some intervals $[0,\pi_\ell]$ and $[\pi_r,1]$ with $0 < \pi_\ell < \hat{\pi} < \pi_r < 1$; $\underline{m} > m$ on $]\pi_\ell,\pi_r[$; $\underline{m}$ has a continuous first derivative on $[0,1]$ and a continuous second derivative on $[1,0] - \{\pi_\ell,\pi_r\}$.[34] Define $\underline{G}(\pi) = G(\pi,\underline{m}(\pi))$. Next, set

$$c_3 = \max_\pi \underline{G}(\pi), \qquad c_4 = \sup_{\pi \in [\pi_\ell,\pi_r] - \{\hat{\pi}\}} \left[f(\pi)\,\underline{G}(\pi) + (\pi - \tilde{\pi})\,\underline{G}'(\pi) \right]$$

and

$$c_5 = \Delta\beta^2 \min_{\pi_\ell \le \pi \le \pi_r} \frac{\pi^2\,(1-\pi)^2\,\underline{m}''(\pi)}{2\,\beta(\pi)}\,.$$

The constants c_3 and c_5 are clearly positive. As to c_4, there is at least one belief $\pi^\dagger$ in $]\pi_\ell,\pi_r[$ such that $(\pi^\dagger - \tilde{\pi})\,\underline{G}'(\pi^\dagger) \ge 0$, hence $c_4 \ge f(\pi^\dagger)\,\underline{G}(\pi^\dagger) > 0$. Moreover, c_4 is finite since $\underline{G}$ has finite one-sided derivatives at $\hat{\pi}$.

[34] For example, define $\phi(\pi) = (\pi - \pi_\ell)^2(\pi - \pi_r)^2$ on $]\pi_\ell,\pi_r[$ and $\phi(\pi) = 0$ everywhere else; then $\underline{m}(\pi) = [1 + \delta\,\phi(\pi)]\,m(\pi)$ will have the desired properties for $\delta > 0$ sufficiently small.

Now let $r \geq 0$, $\Lambda \geq 0$ and $\sigma > 0$ be such that $c_3 r + c_4 \Lambda \leq c_5/\sigma^2$. By construction, this implies that $\underline{m}$ is a subsolution of (G.4) both to the left and to the right of the central ray $\mathcal{R}_c$. As $\hat{q} = q_c$, this ray is vertical at $\pi = \hat{\pi}$. Arguing as in the proof of Lemma G.4, we see that the right-hand side of the ODE is regular with respect to $\underline{m}$ and $\overline{m}$ on each closed interval contained in $]0, \hat{\pi}]$ or $[\hat{\pi}, 1[$, where it is understood that the appropriate one-sided limit is used to calculate the right-hand side of the ODE at $\hat{\pi}$. The result thus follows from Lemma G.2 and Corollary G.1. ∎

H Numerical Simulations

The adjusted value function can be calculated approximately as a numerical solution to a two-point boundary value problem, namely the ODE (19) subject to the boundary conditions $v^*(0) = m(0)$ and $v^*(1) = m(1)$. We used the method of *relaxation*[35] to do this. Beliefs were discretised with a step size of 10^{-3}, decreasing to 10^{-5} around the confounding belief. The iterative procedure was deemed to have converged when the maximum pointwise difference between successive approximations to the value function and its first derivative were less than 0.0001%. Convergence was quite rapid, varying from 5 iterations for a high discount rate without switching, to 18 iterations for a low discount rate with an intermediate switching intensity close to the critical level. The procedure was implemented on a VAX minicomputer under VMS v5.4. Each iteration took approximately 19 seconds of CPU time, so the numerical solutions each took between only 1.5 and 6 minutes to calculate.

Given a numerical approximation to the adjusted value function, the optimal policy correspondence immediately yields an approximately optimal policy function. To generate sample paths of posterior beliefs and optimal quantities, we first chose an initial state and an initial belief. One iteration then consisted of the following steps: (a) calculate the optimal quantity given the current belief (using the above numerical results); (b) introduce a shock; (c) update the belief using equation (4) in its discrete form, namely

$$\begin{aligned} \delta\pi_t = \lambda(\pi_t)\,\delta t \quad &+ \quad \sigma^{-2}\pi_t(1-\pi_t)(k_t-\pi_t)(\Delta\alpha - \Delta\beta\, q_t)^2\,\delta t \\ &+ \quad \sigma^{-1}\pi_t(1-\pi_t)(\Delta\alpha - \Delta\beta\, q_t)\,\delta Z_t\,; \end{aligned}$$

(d) update the state if required (depending on the transition probabilities λ_0 and λ_1). These four steps are then repeated to generate a succession of beliefs and quantities. State switching was implemented by repeatedly drawing a number from the uniform

[35] See Press *et al.* (1986), Chapter 16.

distribution on the unit interval (all the examples reported in the chapter have $\tilde{\pi} = 0.5$, that is $\lambda_0 = \lambda_1 = \Lambda/2$). If the number drawn is less than $1 - \exp(-\Lambda/2)$, then the state remains unchanged, else it switches. Over a time interval of 100, we expect to see 10 switches for $\Lambda = 0.2$. For other values of Λ, the time interval is 'stretched' accordingly, so for $\Lambda = 0.05$, for example, we expect these 10 switches to occur by the time $t = 400$.

The shocks were generated by repeated draws from the standard normal distribution. For given time increment δt, the shock δZ was taken to be $\sqrt{\delta t}$ times the draw from the standard normal distribution.

In order to maintain a reasonable approximation to the continuous case that we are modelling, we must ensure that each $\delta\pi$ is not so large that the agent's belief can jump to (or past) 0, 1, or $\hat{\pi}$. To achieve this, the time variable was incremented by 0.05 in each discrete period, i.e. $\delta t = 0.05$. (This means that in the cases without state switching there are several hundred iterations, and in those with state switching there are a few thousand.)

References

Aghion, P., Bolton, P., Harris, C. and Jullien, B. (1991): "Optimal Learning by Experimentation", *Review of Economic Studies*, **58**, 621–654.

Bala, V. and Kiefer, N.M. (1990): "Information Investment and Dynamic Market Performance" (CAE Working Paper No. 90-10, Cornell University).

Balvers, R.J. and Cosimano, T.F. (1990): "Actively Learning about Demand and the Dynamics of Price Adjustments", *Economic Journal*, **100**, 882–898.

Balvers, R.J. and Cosimano, T.F. (1993): "Periodic Learning about a Hidden State Variable", *Journal of Economic Dynamics and Control*, **17**, 805–827.

Balvers, R.J. and Cosimano, T.F. (1994): "Inflation Variability and Gradualist Monetary Policy", *Review of Economic Studies*, **61**, 721–738.

Banks, J.S. and Sundaram, R.K. (1992): "Denumerable-armed Bandits", *Econometrica*, **60**, 1071–1096.

Banks, J.S. and Sundaram, R.K. (1994): "Switching Costs and the Gittins Index", *Econometrica*, **62**, 687–694.

Bergemann, D. and Välimäki, J. (1996): "Market Experimentation and Pricing" (Cowles Foundation Discussion Paper No. 1122, Yale University).

Bergemann, D. and Välimäki, J. (1997): "Market Diffusion with Two-Sided Learning", *RAND Journal of Economics*, **28**, 773–795.

Bernfeld, S.R. and Lakshmikantham, V. (1974): *An Introduction to Nonlinear Boundary Value Problems* (New York and London: Academic Press).

Berry, D.A. and Fristedt, B. (1985): *Bandit Problems: Sequential Allocation of Experiments* (New York and London: Chapman & Hall).

Bertocchi, G. and Spagat, M. (1993): "Learning, Experimentation, and Monetary Policy", *Journal of Monetary Economics*, **32**, 169–183.

Bolton, P. and Harris, C. (1993): "Strategic Experimentation" (STICERD Discussion Paper No. TE/93/261, London School of Economics).

Diamond, P.A. (1982): "Wage Determination and Efficiency in Search Equilibrium", *Review of Economic Studies*, **44**, 217–227.

Dixit, A. and Pindyck, R. (1994): *Investment under Uncertainty* (Princeton: Princeton University Press).

Dutta, P.K. (1991): "What Do Discounted Optima Converge to? A Theory of Discount Rate Asymptotics in Economic Models", *Journal of Economic Theory*, **55**, 64–94.

Easley, D. and Kiefer, N.M. (1988): "Controlling a Stochastic Process with Unknown Parameters", *Econometrica*, **56**, 1045–1064.

Felli, L. and Harris, C. (1996): "Learning, Wage Dynamics and Firm-Specific Human Capital", *Journal of Political Economy*, **104**, 838–868.

Fleming, W.H. and Rishel, R.W. (1975): *Deterministic and Stochastic Control* (New York: Springer-Verlag).

Furusawa, T. (1994): "Trial and Error in R&D" (mimeo, University of Wisconsin-Madison).

Gittins, J.C. (1979): "Bandit Processes and Dynamic Allocation Indices", *Journal of the Royal Statistical Society B*, **41**, 148–164.

Gittins, J.C. (1989): *Multi-armed Bandit Allocation Indices* (Chapter 3. Chichester: Wiley).

Gittins, J.C. and Jones, D.M. (1974): "A Dynamic Allocation Index for the Sequential Design of Experiments", in *Progress in Statistics*, (J.Gani, ed.) pp. 241–266. Amsterdam: North-Holland.

Glazebrook, K.D. (1982): "On a Sufficient Condition for Superprocesses due to Whittle", *Journal of Applied Probability*, **19**, 99–110.

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

Harris, C. (1988): "Dynamic Competition for Market Share: An Undiscounted Model" (working paper, Nuffield College, Oxford).

Jovanovic, B. (1979): "Job Matching and the Theory of Turnover", *Journal of Political Economy*, **87**, 972–990.

Karatzas, I. (1984): "Gittins Indices in the Dynamic Allocation Problem for Diffusion Processes", *Annals of Probability*, **12**, 173–192.

Karlin, S. and Taylor, H.M. (1981): *A Second Course in Stochastic Processes* (New York: Academic Press).

Kiefer, N.M. (1989a): "A Value Function Arising in the Economics of Information", *Journal of Economic Dynamics and Control*, **13**, 201–223.

Kiefer, N.M. (1989b): "A Dynamic Model of Optimal Learning with Obsolescence of Information" (CAE Working Paper No. 89-14, Cornell University; revised 1991).

Krylov, N.V. (1980): *Controlled Diffusion Processes* (New York: Springer-Verlag).

Leach, J.C. and Madhavan, A. (1993): "Price Experimentation and Security Market Structure", *Review of Financial Studies*, **6**, 375–404.

Liptser, R.S. and Shiryayev, A.N. (1977): *Statistics of Random Processes I* (New York: Springer-Verlag).

McLennan, A. (1984): "Price Dispersion and Incomplete Learning in the Long

Run", *Journal of Economic Dynamics and Control*, **7**, 331–347.

Miller, R.A. (1984): "Job Matching and Occupational Choice", *Journal of Political Economy*, **92**, 1086–1120.

Mirman, L.J., Samuelson, L. and Urbano, A. (1993): "Monopoly Experimentation", *International Economic Review*, **34**, 549–564.

Moscarini, G. and Smith, L. (1997): "Wald Revisited: The Optimal Level of Experimentation" (Working Paper, Yale University and MIT).

Nash, P. (1973): "Optimal Allocation of Resources between Research Projects" (Ph.D. thesis, Cambridge University).

Nyarko, Y. and Olson, L. (1996): "Optimal Growth with Unobservable Resources and Learning", *Journal of Economic Behavior and Organization*, **29**, 465–491.

Pissarides, C.A. (1990): *Equilibrium Unemployment Theory* (Oxford: Blackwell).

Prendergast, C. and Stole, L. (1996): "Impetuous Youngsters and Jaded Old-Timers: Acquiring a Reputation for Learning", *Journal of Political Economy*, **104**, 1105–1134.

Prescott, E.C. (1972): "The Multiperiod Control Problem under Uncertainty", *Econometrica*, **40**, 1043–1058.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1986): *Numerical Recipes: the Art of Scientific Computing* (Cambridge and New York: CUP).

Revuz, D. and Yor, M. (1991): *Continuous Martingales and Brownian Motion* (New York: Springer-Verlag).

Roberts, K. and Weitzman, M. (1981): "Funding Criteria for Research, Development, and Exploration Projects", *Econometrica*, **49**, 1261–1288.

Rogers, L.C.G. and Williams, D. (1987): *Diffusions, Markov Processes and Martingales, Vol.2: Itô Calculus* (Chichester: Wiley).

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

Rustichini, A. and Wolinsky, A. (1995): "Learning about Variable Demand in the Long Run", *Journal of Economic Dynamics and Control*, **19**, 1283–1292.

Smith, L. (1992): "Error Persistence, and Experiential versus Observational Learning" (Foerder Discussion Paper, Tel Aviv University).

Trefler, D. (1993): "The Ignorant Monopolist: Optimal Learning with Endogenous Information", *International Economic Review*, **34**, 565–581.

Vega-Redondo, F. (1993): "Industrial Dynamics, Path-Dependence, and Technological Change" (mimeo, Universidad de Alicante).

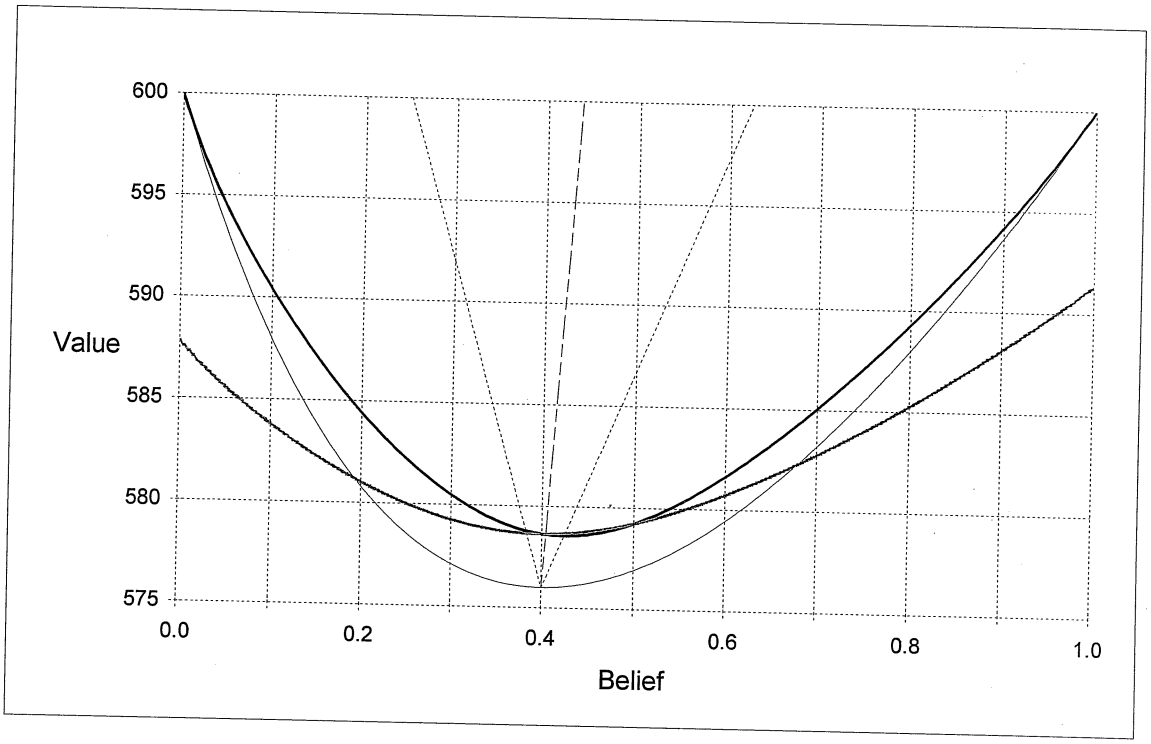
Weitzman, M. (1979): "Optimal Search for the Best Alternative", *Econometrica*, **47**, 641–654.

Whittle, P. (1980): "Multi-armed Bandits and the Gittins Index", *Journal of the Royal Statistical Society B*, **42**, 143–149.

Whittle, P. (1981): "Arm-acquiring Bandits", *Annals of Probability*, **9**, 284–292.

Whittle, P. (1982): *Optimization over Time: Dynamic Programming and Stochastic Control, Vol.I* (Chapter 14. New York: Wiley).

Whittle, P. (1988): "Restless Bandits: Activity Allocation in a Changing World", *Journal of Applied Probability*, **25A**, 287–298.



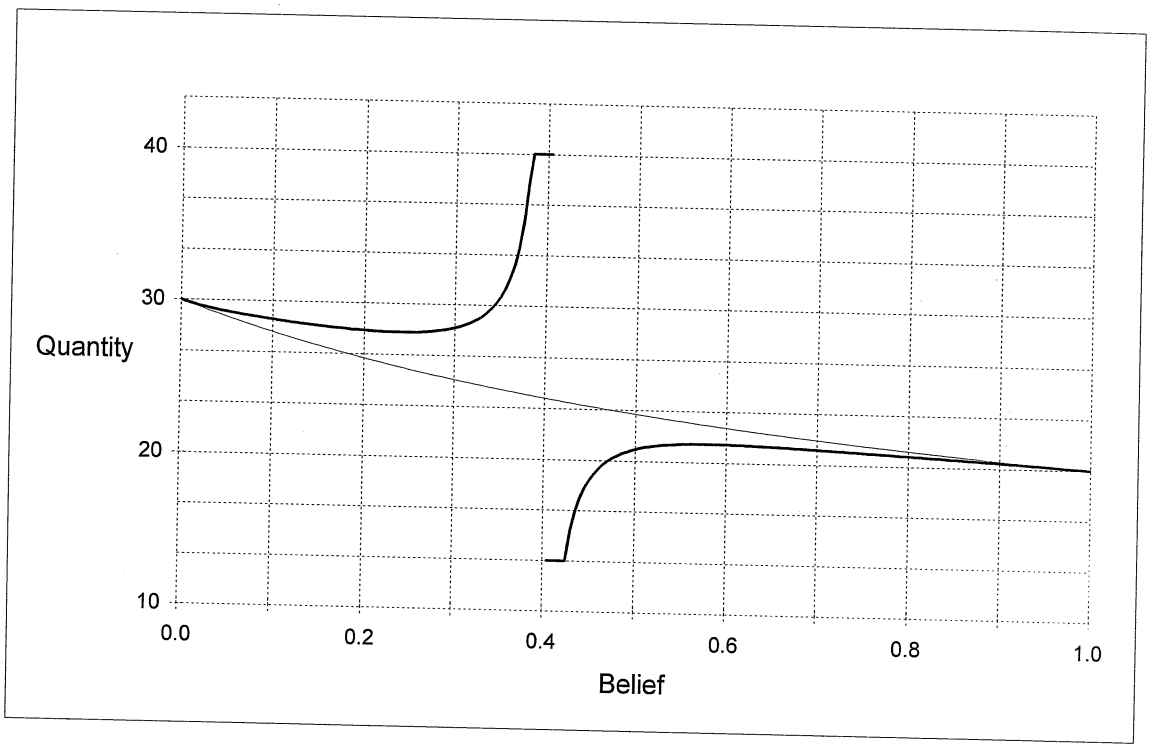
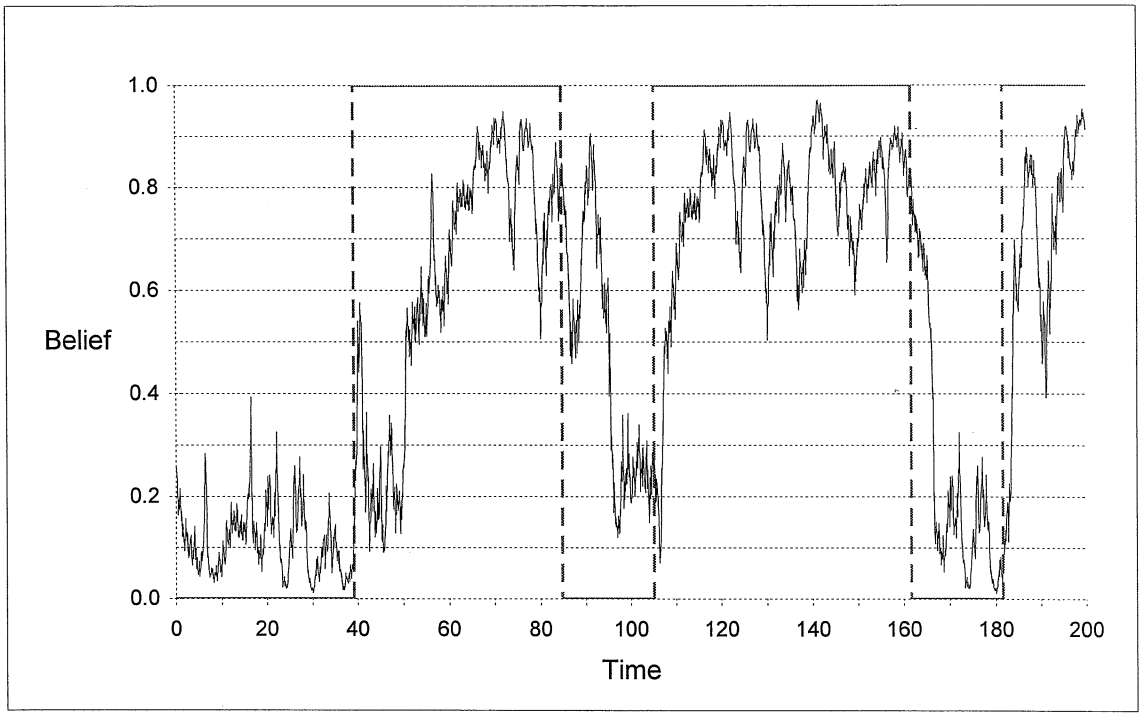


Figure 3.4: Value function & optimal policy for $r = 0.1$, $\Lambda = 0.05$, $\tilde{\pi} = 0.5$; $\hat{\pi} = 0.4$

The bold lines are the adjusted value function v^* and the optimal policy function q^*.
The thin lines are the myopic optimum pay-off m and the myopic policy q^m.
The upper panel shows the three rays $\mathcal{R}_\ell$, $\mathcal{R}_c$ and $\mathcal{R}_r$.
In the upper panel, the value function u^* is plotted as the thick grey line.



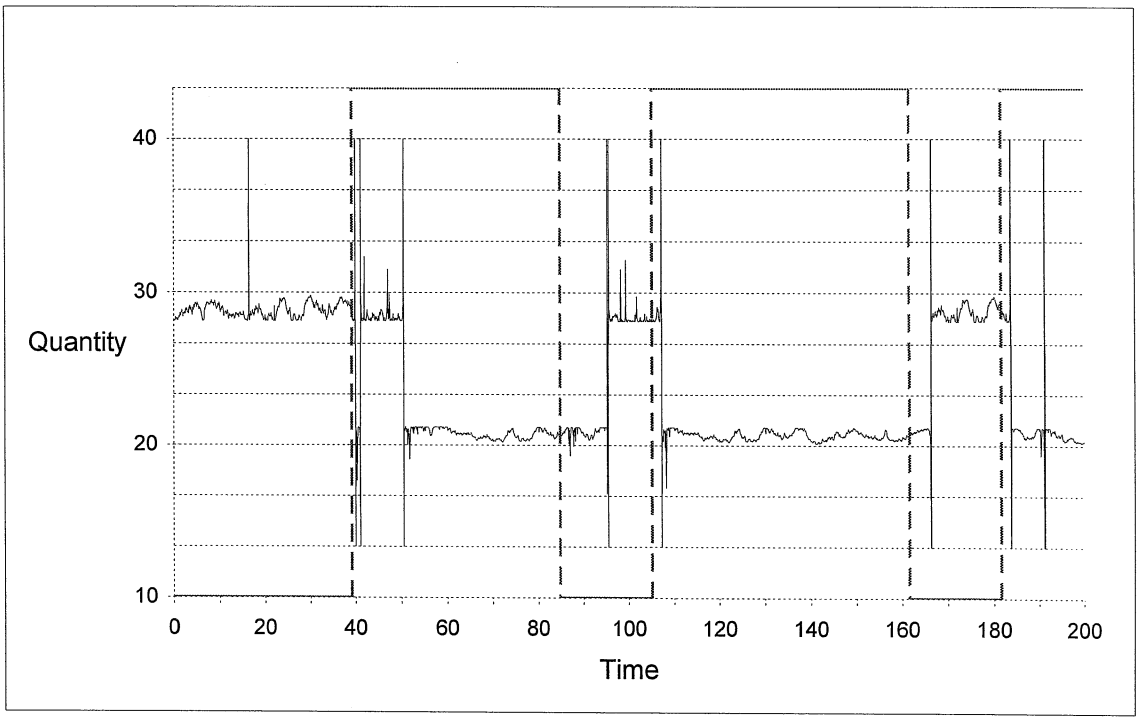
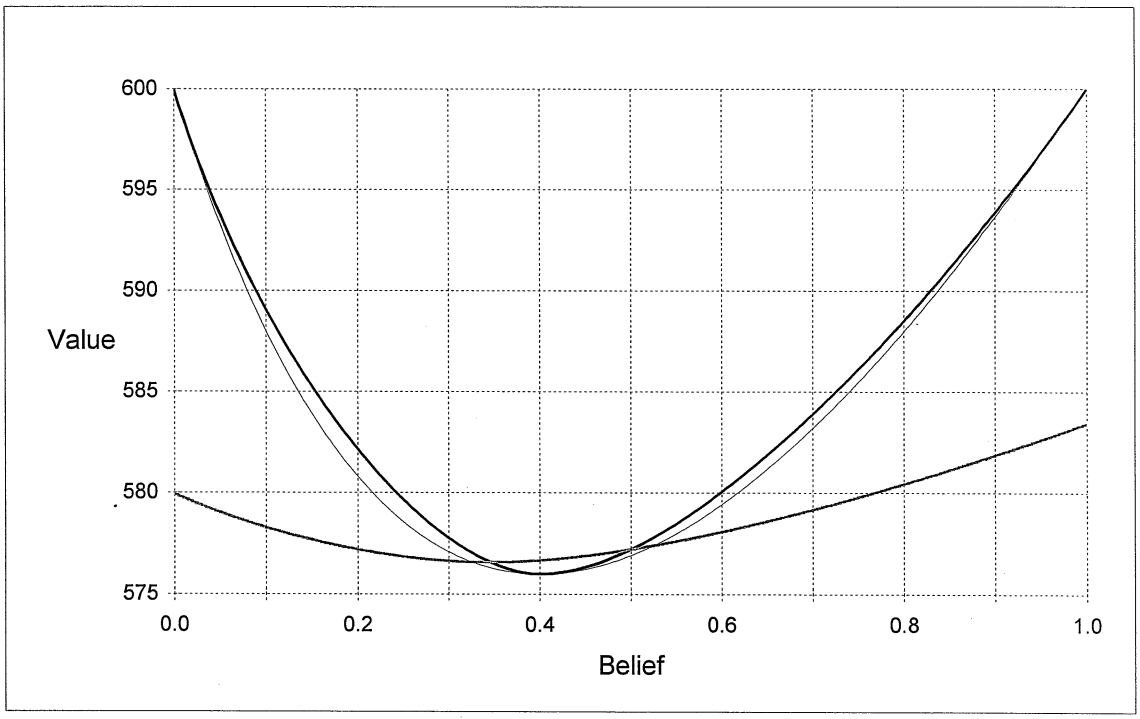


Figure 3.5: Sample paths for $r = 0.1$, $\Lambda = 0.05$, $\tilde{\pi} = 0.5$; $\hat{\pi} = 0.4$

The dashed line indicates the state switches.



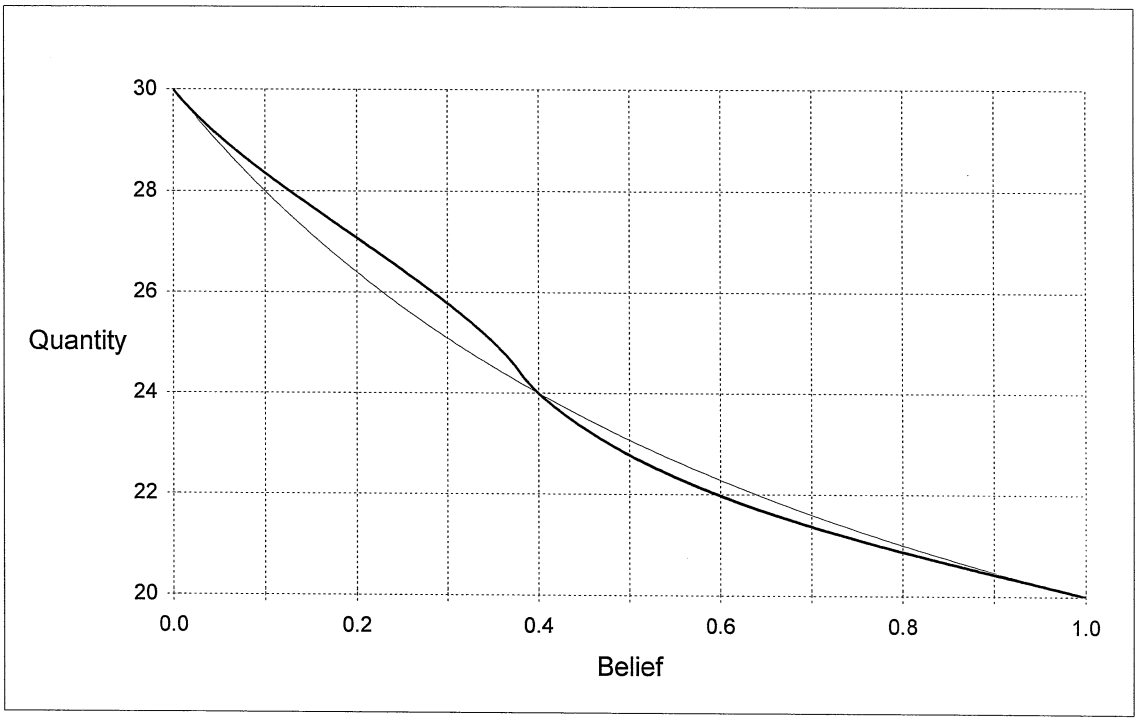
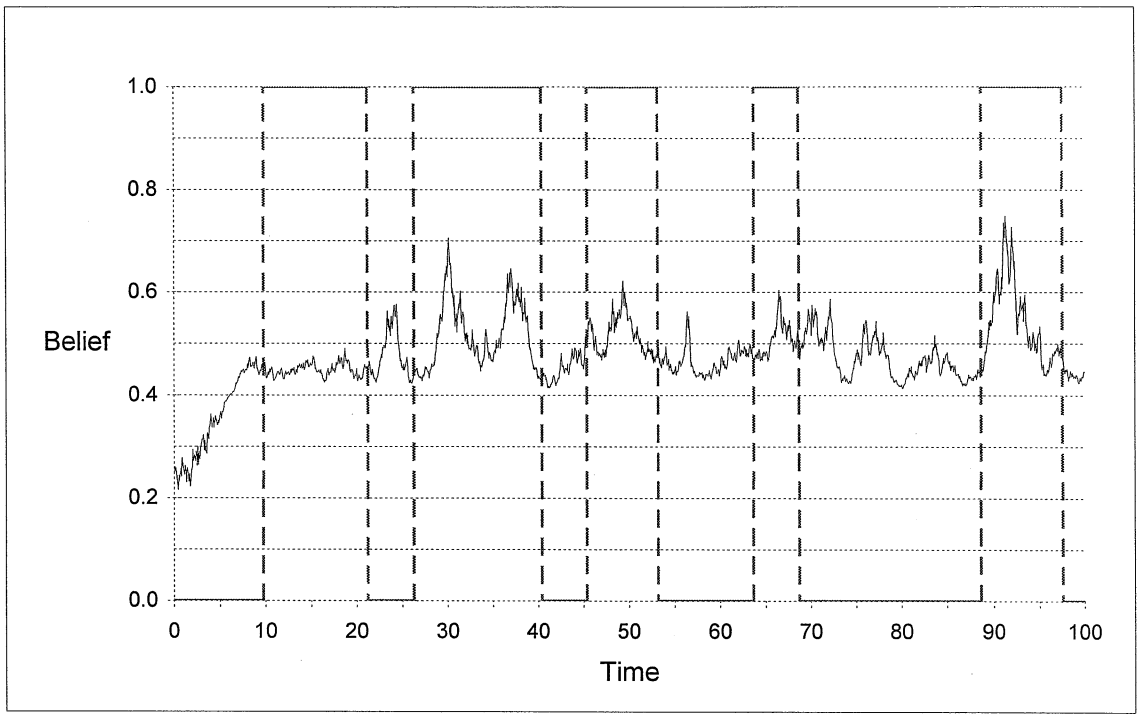


Figure 3.6: Value function & optimal policy for $r = 0.1$, $\Lambda = 0.2$, $\tilde{\pi} = 0.5$; $\hat{\pi} = 0.4$

The bold lines are the adjusted value function v^* and the optimal policy function q^*.
The thin lines are the myopic optimum pay-off m and the myopic policy q^m.
In the upper panel, the value function u^* is plotted as the thick grey line.



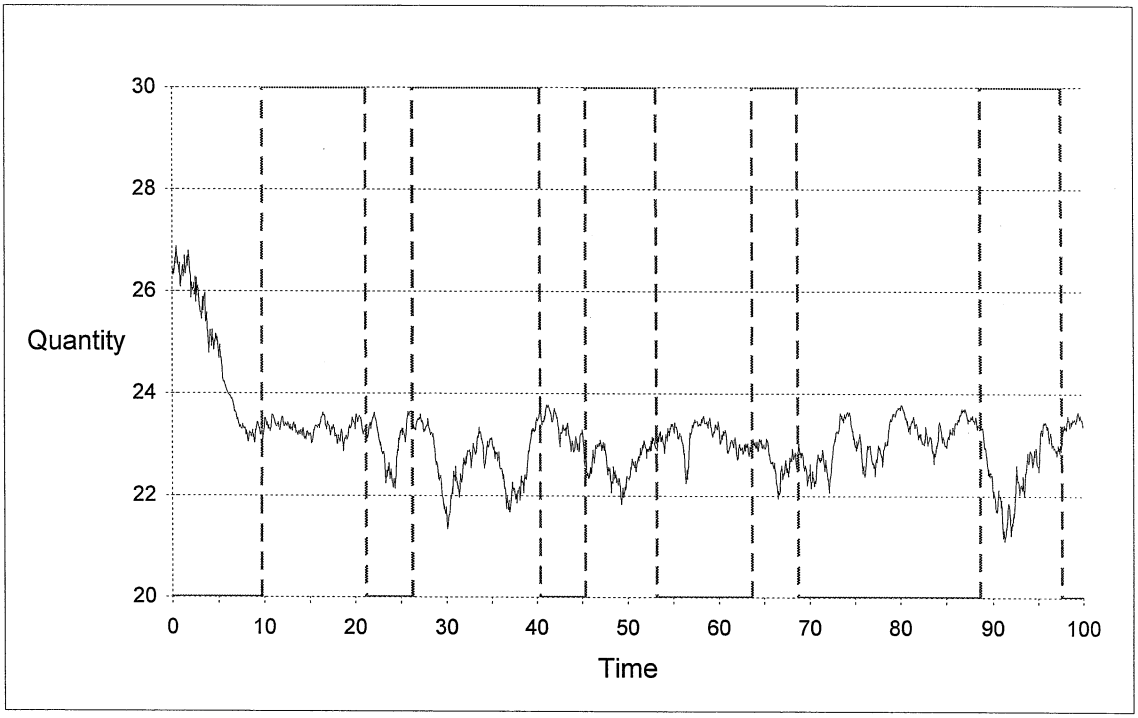
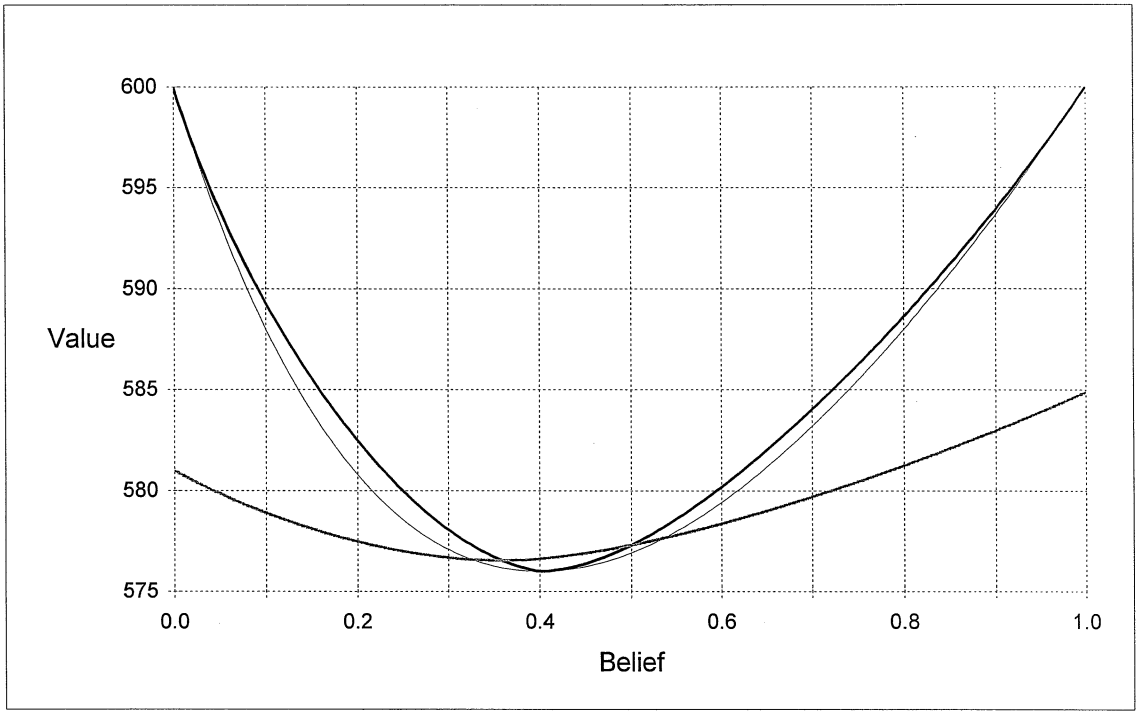


Figure 3.7: Sample paths for $r = 0.1$, $\Lambda = 0.2$, $\tilde{\pi} = 0.5$; $\hat{\pi} = 0.4$

The dashed line indicates the state switches.



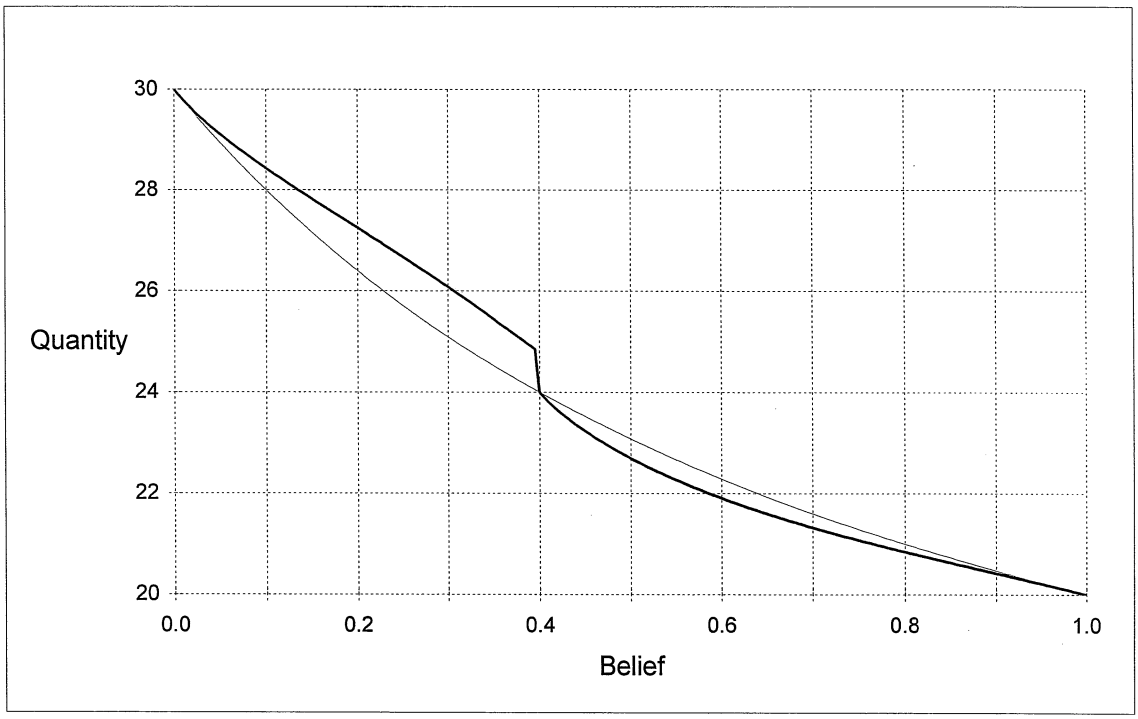
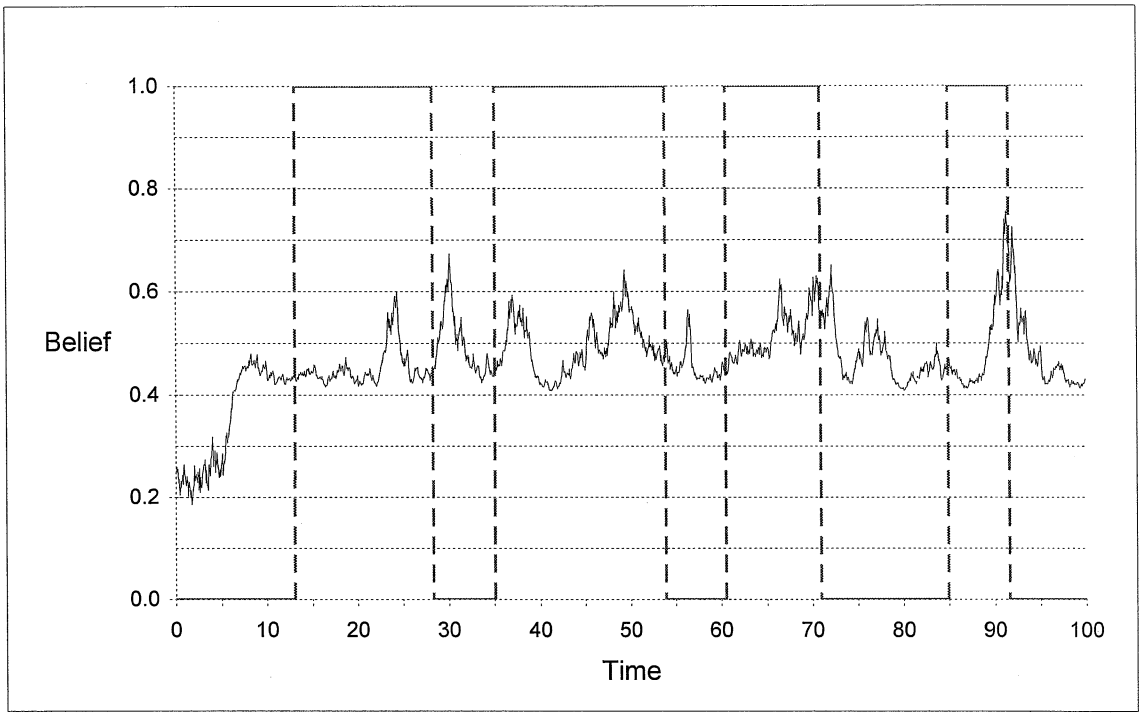


Figure 3.8: Value function & optimal policy for $r = 0.1$, $\Lambda = 0.15$, $\tilde{\pi} = 0.5$; $\hat{\pi} = 0.4$

The bold lines are the adjusted value function v^* and the optimal policy function q^*.
The thin lines are the myopic optimum pay-off m and the myopic policy q^m.
In the upper panel, the value function u^* is plotted as the thick grey line.



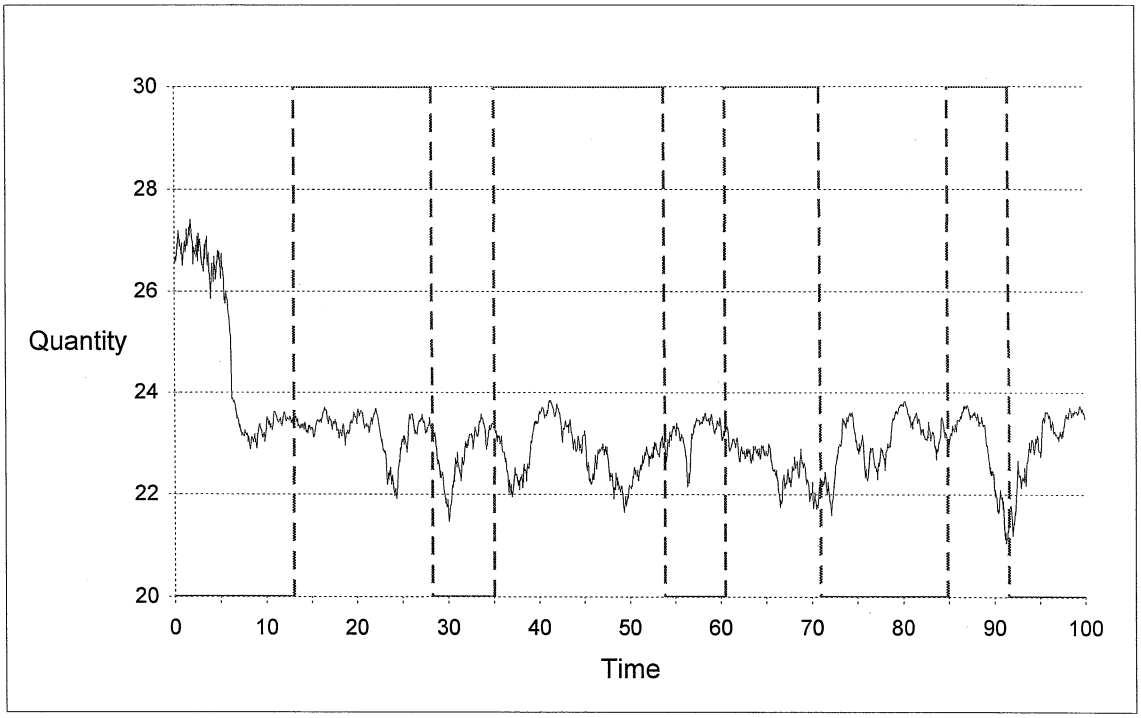
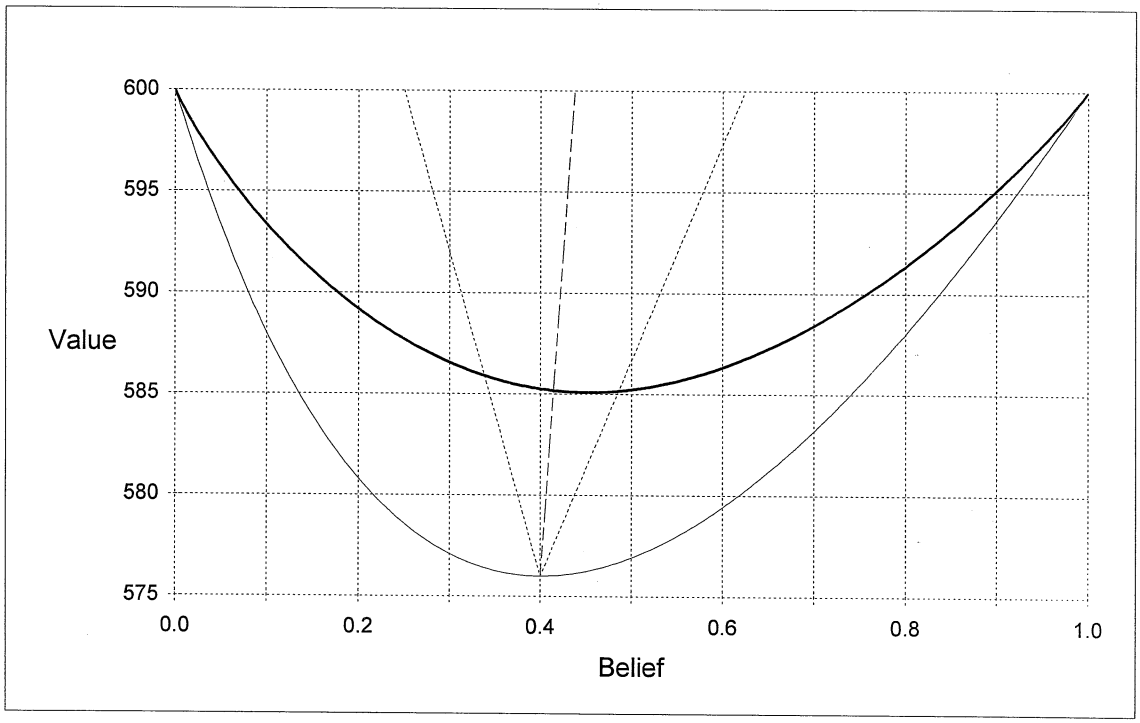


Figure 3.9: Sample paths for $r = 0.1$, $\Lambda = 0.15$, $\tilde{\pi} = 0.5$; $\hat{\pi} = 0.4$

The dashed line indicates the state switches.



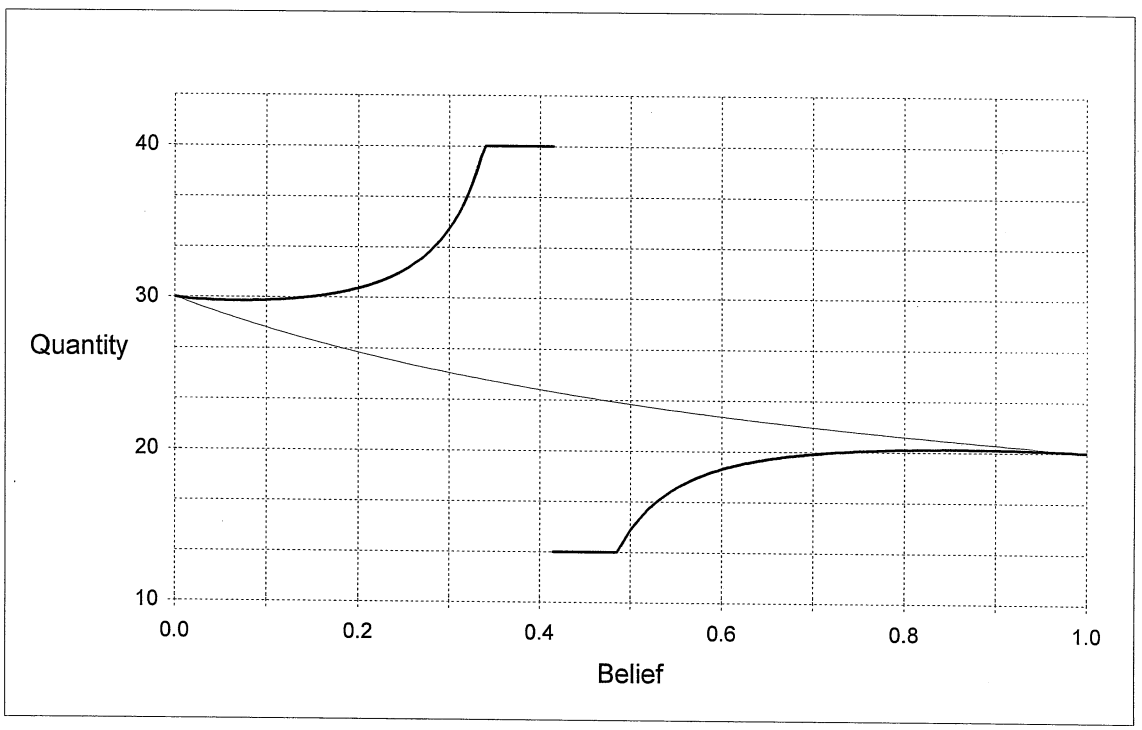
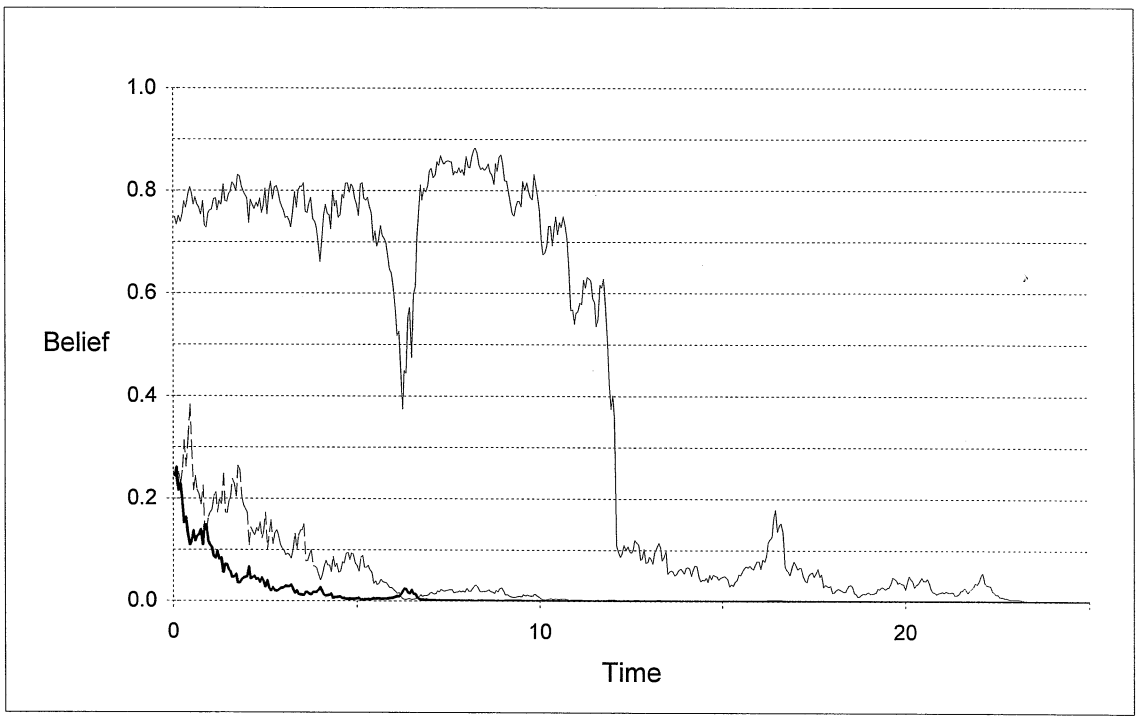


Figure 3.10: Value function & optimal policy for $r = 0.1$, $\Lambda = 0$; $\hat{\pi} = 0.4$

The bold lines are the adjusted value function v^* and the optimal policy function q^*.
The thin lines are the myopic optimum pay-off m and the myopic policy q^m.
The upper panel shows the three rays $\mathcal{R}_\ell$, $\mathcal{R}_c$ and $\mathcal{R}_r$.



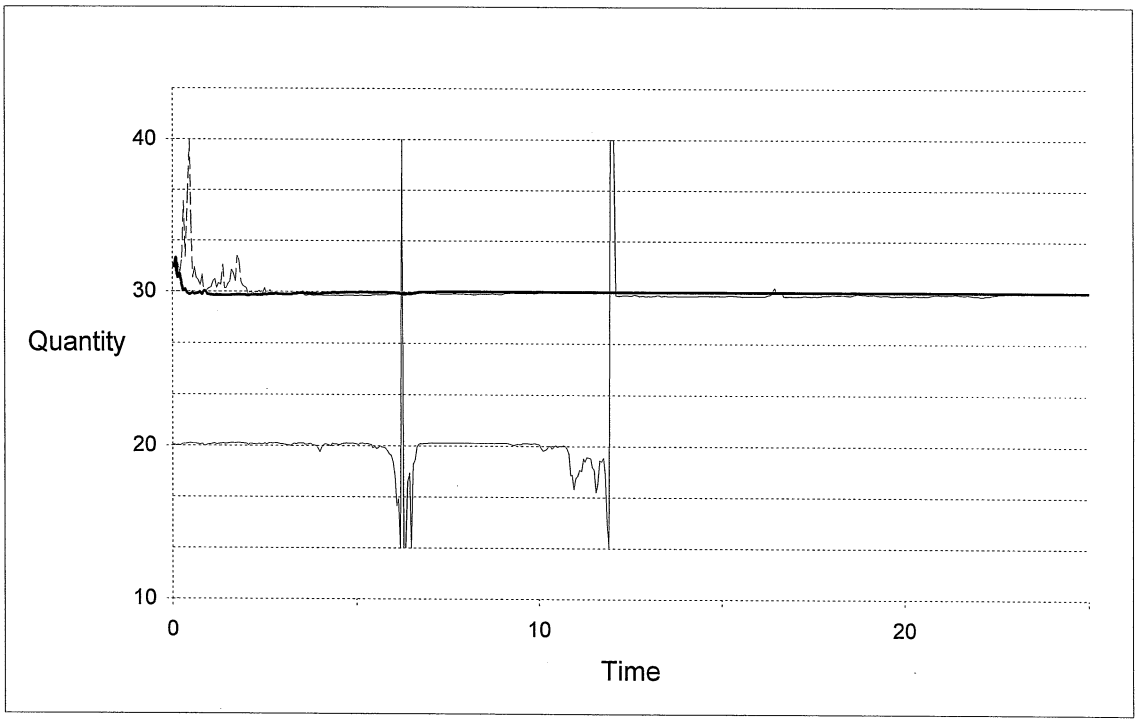
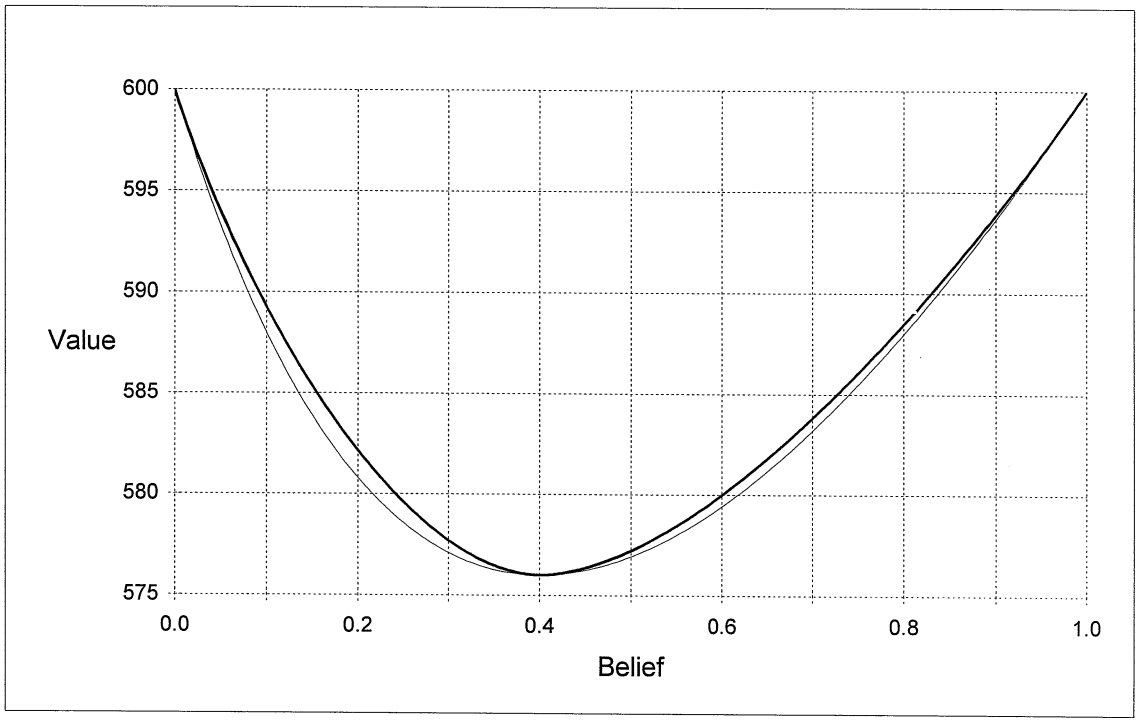


Figure 3.11: Sample paths for $r = 0.1$, $\Lambda = 0$; $\hat{\pi} = 0.4$

The three sample paths are for different prior beliefs or different shocks.



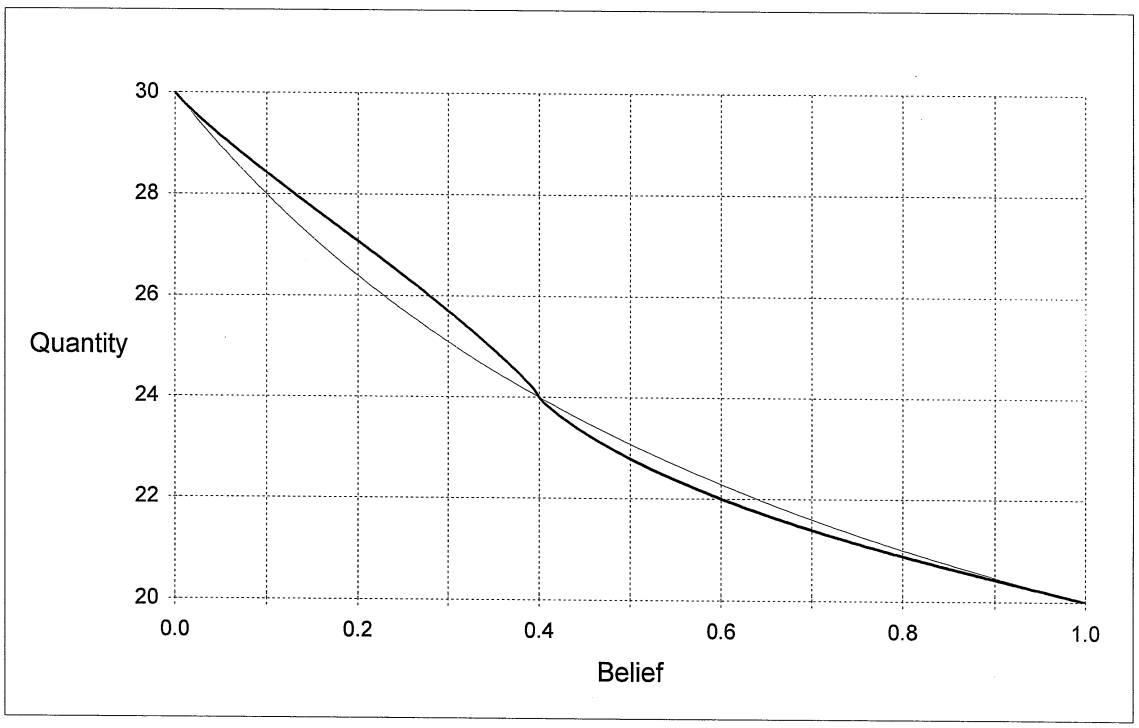
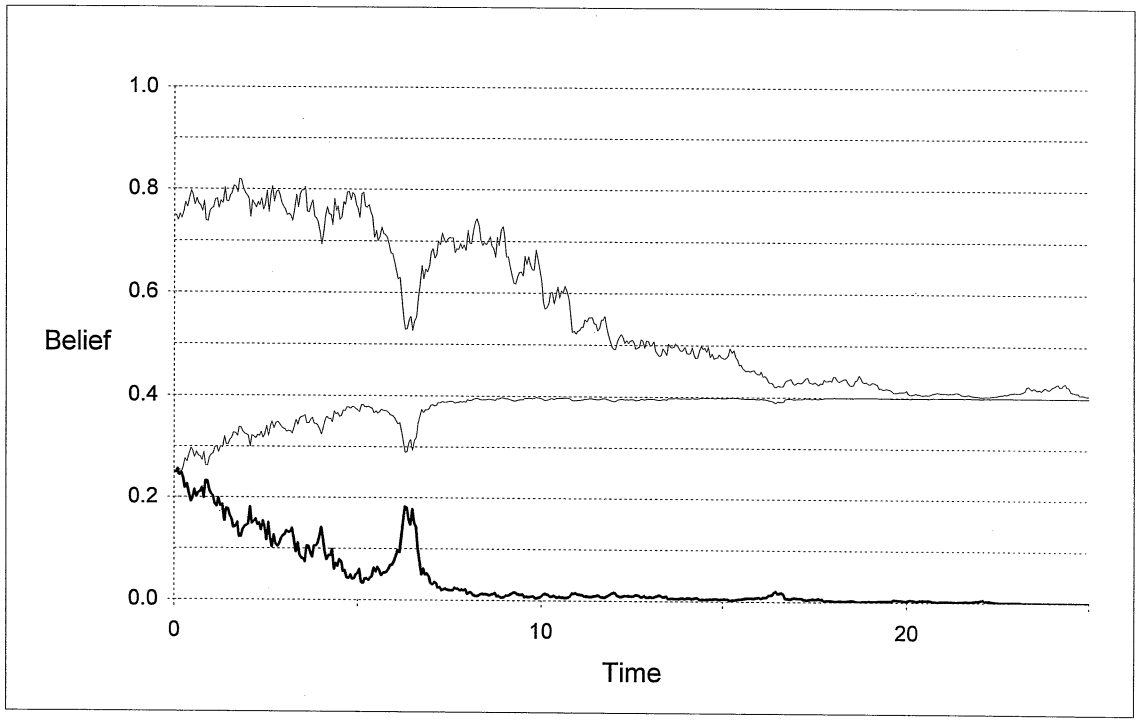


Figure 3.12: Value function & optimal policy for $r = 0.5$, $\Lambda = 0$; $\hat{\pi} = 0.4$

The bold lines are the adjusted value function v^* and the optimal policy function q^*.
The thin lines are the myopic optimum pay-off m and the myopic policy q^m.



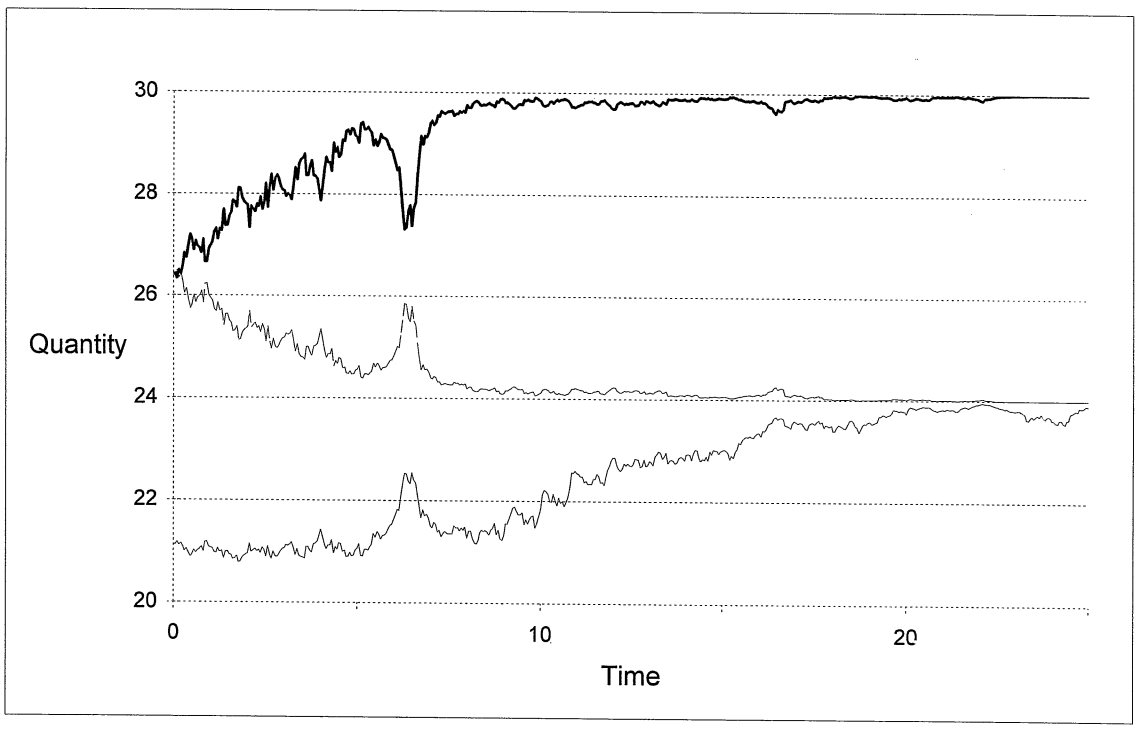


Figure 3.13: Sample paths for $r = 0.5$, $\Lambda = 0$; $\hat{\pi} = 0.4$

The three sample paths are for different prior beliefs or different shocks.