The Mathematics of Bayesian Learning Traps

Simon Loertscher
University of Melbourne*

Andrew McLennan
University of Queensland†

March 28, 2012

PRELIMINARY AND INCOMPLETE.

COMMENTS ARE WELCOME.

Abstract

A Bayesian decision maker does not know which of several parameters is true. In each period she chooses an action a from an open subset of $\mathbb{R}^n$, observes an outcome, and updates her beliefs. There is an action a^* that is uninformative in the sense that when it is chosen all parameters give the same distribution over outcomes, and consequently beliefs do not change. We give conditions under which a policy specifying an action as a function of the current belief can result in a positive probability that the sequence of beliefs converge to a belief at which a^* is chosen, so that learning is asymptotically incomplete. Such a policy can be optimal even when the decision maker is not myopic and values experimentation.

Keywords: Bayesian learning, information cascades, dynamic programming, stochastic optimal control.

1 Introduction

Learning and experimentation are omnipresent in economic life. Firms need to design new products and hire new employees, consumers have to choose menus and restaurants, and politicians have to pick policies in an uncertain world. A natural and relevant

*Email: simonl@unimelb.edu.au.

†Email: a.mclennan@economics.uq.edu.au. McLennan's work was funded in part by Australian Research Council grant DP0773324. A collaboration with Mark Feldman was an important precursor to this project; his insights and influence are gratefully acknowledged.

question is under which conditions a decision maker will eventually learn, or fail to learn, the truth.

This paper describes a fairly general and robust model of "learning traps," by which we mean situations in which an agent, or a group of agents, persistently choose actions that are both suboptimal (for a fully informed agent) and uninformative, because the costs of experimentation exceed the expected benefits. A Bayesian decision maker who is uncertain about an underlying parameter begins each period with a belief, chooses an action, and observes an outcome. There is a known relationship between the parameter, the action, and the distribution of observations, and the belief at the beginning of the next period is given by Bayesian updating. There is an action a^* that is uninformative in the sense that when it is chosen when it is optimal to do so the distribution over outcomes is the same for all parameters, and consequently beliefs do not change.

The mechanism that leads to asymptotic incomplete learning in our setup is roughly as follows. We suppose that there is a policy dictating the choice of action as a continuous function of the current belief, and this policy dictates that a^* is chosen at a critical belief ω^*. Thus the amount of experimentation is small when the prior belief is near ω^*, which makes it hard for the posterior to be much farther away from the critical belief than the prior. At the same time, there can be a significant probability that the posterior is much closer to the critical belief than the prior, resulting in slower learning. We show that policies that allow such asymptotically incomplete learning can be optimal even when the decision maker is not myopic and values experimentation.

Some of the oldest literature of on learning traps concerns multiarmed bandits (Gittens and Jones (1974), Berry and Fristedt (1985), and references therein). Rothschild (1974) introduced this topic to economics, with subsequent contributions by McLennan (1984), Kihlstrom et al. (1984), Easley and Kiefer (1988), Aghion et al. (1991), Smith and Sørensen (2000), and others. Multiagent environments of this sort are studied by Banerjee (1992) and Bikhchandani et al. (1992), and numerous subsequent papers. Models with this kind of feature have been applied to political economy (e.g., Piketty (1995)) regulation (e.g., Laslier et al. (2003) and Berentsen et al. (2008)), pricing in industrial organization (e.g. Harrison et al. (2011)) and law and economics (e.g., Baker and Mezzetti (2011)). See Smith and Sørensen (2011) for a recent survey and summary.

From a mathematical point of view there are several reasons one might fail to experiment. If the space of possible choices is discrete, as in the bandit literature (e.g., Gittens and Jones (1974), Rothschild (1974), and Banks and Sundaram (1992)) and the literature on information cascades (Banerjee (1992) and Bikhchandani et al. (1992)), there may be a positive lower bound on the costs of experimentation in a single period. In addition,

there may be switching costs (e.g., Banks and Sundaram (1994)) which loom large in many labor market applications. When the space of actions has an uninformative action on its boundary, the expected loss in the current period, and the amount of information acquired resulting from moving away from that point will typically be proportional to the distance moved (see Rothschild and Stiglitz (1984) for one formalization of this notion) and in many setups it is easy to see that not experimenting is optimal when the future is heavily discounted.

Finally, there is possibility that the relevant space of beliefs is (homeomorphic to) an open subset of a Euclidean space, and that there is a positive probability that optimal behavior will induce a sequence of beliefs that converges to a critical belief at which the optimal action is uninformative. McLennan (1984) considered the case in which the space of beliefs is one dimensional, because the unknown parameter has two possible values. It is easy to construct examples in which there is an action that is uninformative, in the sense that the distribution of outcomes does not depend on the unknown parameter when it is chosen, and this action is chosen by the myopically optimal policy in response to a certain critical belief. It can easily happen that the sequence of beliefs cannot go from one side of the critical belief to the other, because the amount of experimentation is never sufficient. McLennan (1984) showed that the optimal policy can have this property even when the discount factor is positive. That is, even when the decision maker cares about the future, it can be optimal to behave in a way that sometimes results in the true value of the parameter remaining unknown in the limit.

The remainder has the following organization. In Section 2 we present a general model of learning as a controlled Markov process: the state of the system is the belief about an unknown parameter, there is a stationary policy function mapping the belief to a space of actions, an outcome is observed, and Bayesian updating gives rise to a posterior belief, which is the next period's state. The first main result describes conditions on the policy function under which there is a positive probability of incomplete learning in the sense that the sequence of beliefs converges to a belief at which the action prescribed by the policy is uninformative. Intuitively, the policy function will induce a positive probability of convergence when the logarithm of the distance from the current belief to the potential limit is a supermartingale when the belief is near the potential limit. Section 3 develops a version of the law of large numbers that formalizes this intuition, and uses it to prove the first main result. Although we have not found our particular version of the law of large numbers elsewhere, the subject is certainly well explored, and Appendix 1 of Ellison and Fudenberg (1995) and Appendix C of Smith and Sørensen (2000) present similar results, so there seems to be little reason to think that this aspect

of the work has a high degree of novelty.

The second main result is described informally at the end of Section 2, then stated and proved in Section 5. The criterion for a positive probability of convergence given by the first result requires that the policy map the critical belief to the uninformative action, and that the derivative of the policy function satisfy certain conditions. We wish to show that when these conditions are satisfied by the optimal myopic policy, they are also satisfied by the optimal policy of a decision maker whose discount factor is positive and small. In McLennan (1984) this aspect of the matter was handled in a concrete and ad hoc manner, but here we give a general result providing conditions under which the optimal policy varies continuously in the C^1 topology (that is, both the policy and its first derivative vary continuously) as we vary the discount factor near zero. The most important hypotheses of this result are that the optimal myopic policy satisfies the second order conditions strictly, and that the operator passing from a C^2 value function for the value of tomorrow's state to the expected valued of tomorrow's state, as a function of today's state and action, is continuous relative to the C^2 topology. This result seems to have considerable independent interest, and to the best of our knowledge it is novel, but of course dynamic programming is also very well studied, so it would not be surprising if these methods had already been developed elsewhere.

Our results require that the space of beliefs and the space of actions have the same dimension. (Among other things, this allows the theory of the topological degree to be used to show that policy functions that are near the myopic policy also map some belief to the uninformative action.) The conditions identified by the first main result require that a certain quantity be negative at every point in a sphere of one lower dimension. As we mentioned above, when the spaces of beliefs and actions are one dimensional it is obvious that the conditions of the first main result can be satisfied, but for higher dimensions the issue is, at this point, unsettled. A related issue is the extent to which adopting a policy of less aggressive experimentation increases the likelihood of incomplete learning. Section 6 explains how to compute the relevant quantities, and considers a pertinent but inconclusive example.

Section 7 discusses possible generalizations and extensions, thereby concluding the paper.

2 Model

We first declare notation and conventions concerning probability. For any measurable space S, let $\Delta(S)$ be the set of probability measures on S, and for $s \in S$ let δ_s be the

Dirac measure of s, i.e., the element of $\Delta(S)$ that assigns all probability to s. Whenever S is a topological space it has the Borel σ-algebra, and $\Delta(S)$ is endowed with the weak* topology; recall that this is the weakest topology such that $\sigma \mapsto \int_S f\, d\sigma$ is a continuous function from $\Delta(S)$ to $\mathbb{R}$ whenever $f : S \to \mathbb{R}$ is continuous and bounded. If S is finite, elements of $\Delta(S)$ are treated as functions from S to [0,1], and $\Delta^\circ(S)$ is the set of $\sigma \in \Delta(S)$ such that $\sigma(s) > 0$ for all s.

Let Θ be a finite set of possible values of a *parameter* $\tilde{\theta}$ that the decision maker may learn about over time. Then $\Omega = \Delta(\Theta)$ is the set of possible *beliefs* concerning $\tilde{\theta}$. This notation reflects a perspective in which the decision maker's current belief is the state of the system.

In each period the decision maker chooses an *action* from an open set $A \subset \mathbb{R}^n$ and observes an *outcome* that is an element of a finite set Y. For each $\theta \in \Theta$ there is a function

$$q_\theta : A \to \Delta^\circ(Y)$$

specifying a probability distribution over outcomes for each action. We always assume that for each θ and y, $q_\theta(y|\cdot) : A \to (0,1)$ is continuous, and for the most part this function will be C^1. If $\omega \in \Omega$ is a prior belief, action a is chosen, and outcome y is observed, then the Bayesian posterior belief is $\beta(\omega, y, a) \in \Omega$ with components given by Bayes rule:

$$\beta_\theta(\omega, a, y) = \frac{\omega_\theta q_\theta(y|a)}{\sum_{\theta' \in \Theta} \omega_{\theta'} q_{\theta'}(y|a)}.$$

We study stochastic processes $\{\tilde{\omega}^t\}$, $\{\tilde{a}^t\}$, and $\{\tilde{y}^t\}$ for dates $t \geq 0$ with $\tilde{\omega}_{t+1} = \beta(\tilde{\omega}_t, \tilde{a}_t, \tilde{y}_t)$ almost surely for all t. We say that *learning is asymptotically incomplete* if $\{\tilde{\omega}_t\}$ does not converge to a point in $\{\, \delta_\theta : \theta \in \Theta \,\}$. When the prior belief is ω and action a is chosen, there is a distribution $B(\omega, a) \in \Delta(\Omega)$ of the posterior belief given by

$$B(\omega, a)(E) = \sum_\theta \omega_\theta q_\theta(a)\big(\{\, y \in Y : \beta(\omega, a, y) \in E \,\}\big).$$

A general property of Bayesian updating is that the expectation of the posterior is the prior:

$$\int \omega' dB(\omega, a) = \omega.$$

It follows that $\{\tilde{\omega}_t\}$ is a martingale: conditional on $\tilde{\omega}_t$ (and regardless of $\tilde{a}_t$) the expectation of $\tilde{\omega}_{t+1}$ is $\tilde{\omega}_t$. Consequently the martingale convergence theorem implies that $\{\tilde{\omega}_t\}$ converges almost surely, so there can be a positive probability of asymptotically incomplete learning only if there is a positive probability of convergence to a belief at which $\tilde{\theta}$ is not known with certainty.

For the most part we will assume that the choice of actions is governed by a stationary policy function

$$p : \Omega \to A.$$

That is, for all t, it is almost surely the case that $\tilde{a}_t = p(\tilde{\omega}_t)$. In this case $B(\tilde{\omega}_t, p(\tilde{\omega}_t))$ is the distribution of $\tilde{\omega}_{t+1}$ conditional on $\tilde{\omega}_t$, so we may think of $\tilde{\omega}_0, \tilde{\omega}_1, \tilde{\omega}_2, \ldots$ as a stationary Markov process.

We are primarily concerned with the possibility that there is a positive probability that learning is asymptotically incomplete because the sequence $\tilde{\omega}_t$ converges to an $\omega^* \in \Delta^\circ(\Theta)$. (Although we will eventually consider the possibility of convergence to a ω^* whose support is neither a singleton nor all of Θ, for the most part, and for the time being, we will only consider the possibility that ω^* has full support.) An action a^* is *uninformative* if $q_\theta(a^*) = q_{\theta'}(a^*)$ for all $\theta, \theta' \in \Theta$. If a^* is uninformative, then $\beta(\omega, a^*, y) = \omega$ for all ω and y, so that $B(\omega, a^*) = \delta_\omega$. If $p(\omega^*)$ is uninformative and $\tilde{\omega}_t = \omega^*$, then it is almost surely the case that $\tilde{\omega}_s = \omega^*$ for all $s \geq t$. There are examples in which this situation arises with positive probability, but they are rather special.

The more interesting possibility is that there is a positive probability that $\tilde{\omega}_t \to \omega^*$ even though $\tilde{\omega}_t \neq \omega^*$ for all t almost surely. If p is continuous, then the function $\omega \mapsto B(\omega, p(\omega))$ is continuous, so there cannot be a positive probability of convergence to an $\omega^* \in \Delta^\circ(\Theta)$ unless $p(\omega^*) = a^*$ is uninformative. So suppose that $p(\omega^*)$ is uninformative.

There are now two main questions:

1) Under what conditions on p will there be a positive probability that $\tilde{\omega}_t \to \omega^*$?

2) When can we be certain that these conditions will be satisfied by the optimal policy of a decision maker who maximizes the expectation of a sum $\sum_{t=0}^{\infty} \delta^t R(a_t, y_t)$ of discounted rewards when δ is positive and sufficiently small?

Remark: Consider the possibility that, as in McLennan (1984), Ω is one dimensional, because $\Theta = \{\theta_1, \theta_2\}$ has two elements. Let A be an open subset of $\mathbb{R}$, let $a^* \in A$ be an action that is uninformative, and let ω^* be a belief such that $p(\omega^*) = a^*$. Suppose that $\tilde{\omega}_0 = \omega_0$ almost surely, where ω_0 is between δ_{θ_1} and ω^*. It can easily happen that for all t, $\tilde{\omega}_t$ is almost surely between δ_{θ_1} and ω^* because the action prescribed by p is never informative enough to move the belief out of this interval. If this is the case, then the sequence of beliefs will almost surely converge to either δ_{θ_1} or ω^*, and, using the fact that $\{\tilde{\omega}_t\}$ is a Martingale, one can easily compute the probabilities of these limits conditional on the actual parameter. Now suppose that the policy p_δ maximizes the expectation of a sum $\sum_{t=0}^{\infty} \delta^t r(a_t, y_t)$ of discounted rewards, where $0 \leq \delta < 1$. It is not hard to construct

examples in which p_0 has all the features described above. McLennan (1984) showed that it can happen that for all δ in some interval $[0, \overline{\delta})$, p_δ also has all of these features.

This paper presents a more general and robust mechanism leading to asymptotic incomplete learning. It is not limited to the case of two possible parameters, nor does it depend on it being impossible to learn that the parameter has certain values. In addition, sufficient conditions for positive probability of asymptotic incomplete learning can easily be checked using the tools of calculus. We now describe the main features of the mechanism informally and intuitively.

Of course $\tilde{\omega}_t \to \omega^*$ if and only if $\ln \|\tilde{\omega}_t - \omega^*\| \to -\infty$. The guiding intuition is that the reasoning underlying the law of large numbers can be applied to the sum

$$\ln \|\tilde{\omega}_t - \omega^*\| - \ln \|\tilde{\omega}_0 - \omega^*\| = \sum_{s=0}^{t-1} \ln \frac{\|\tilde{\omega}_{s+1} - \omega^*\|}{\|\tilde{\omega}_s - \omega^*\|}.$$

In order to do this we will need to provide sufficient information concerning

$$\mathbf{E}\left(\ln \frac{\|\tilde{\omega}_{t+1} - \omega^*\|}{\|\tilde{\omega}_t - \omega^*\|} \Big| \tilde{\omega}_t \right)$$

when $\tilde{\omega}_t$ is close to ω^*. When Ω is one dimensional and the process $\{\tilde{\omega}_t\}$ is confined one of the two intervals determined by ω^*, this expectation is always negative because the process is a martingale and the logarithm function is concave. Thus the mechanism developed here encompasses the phenomenon identified by McLennan (1984).

Let $H = \{\, \omega \in \mathbb{R}^\Theta : \sum_\theta \omega_\theta = 1 \,\}$ be the hyperplane in $\mathbb{R}^\Theta$ that contains Ω, and let

$$S = \{\, \sigma \in \mathbb{R}^\Theta : \|\sigma\| = 1 \text{ and } \sum_\theta \sigma_\theta = 0 \,\}$$

be the unit sphere in the hyperplane through the origin parallel to H. Any $\omega \in H$ has a representation of the form $\omega = \omega^* + r\sigma$ where $\sigma \in S$ and $r \geq 0$, and this representation is unique if $\omega \neq \omega^*$.

Set

$$\Xi = \{\, (\sigma, r) \in S \times (0, \infty) : \omega^* + r\sigma \in \Omega \,\} \quad \text{and} \quad \overline{\Xi} = \Xi \cup (S \times \{0\}).$$

Let $\kappa : \overline{\Xi} \times Y \to \Omega$ be the function

$$\kappa(\sigma, r, y) = \beta(\omega^* + r\sigma, p(\omega^* + r\sigma), y).$$

From this point forward we assume that p and the functions $q_\theta(y|\cdot)$ are C^1. Elementary calculus implies that $\frac{\partial \kappa}{\partial t}$ is a well defined continuous function from $\overline{\Xi}$ to the hyperplane through the origin parallel to H. We define $\psi : \overline{\Xi} \times Y \to H$ by setting

$$\psi(\sigma, r, y) = \begin{cases} \omega^* + \frac{1}{r}(\kappa(\sigma, r, y) - \omega^*), & (\sigma, r) \in \Xi, \\ \omega^* + \frac{\partial \kappa}{\partial r}(\sigma, 0, y), & r = 0. \end{cases}$$

Clearly ψ is continuous at every point of Ξ, and the restriction of ψ to $S \times \{0\}$ is also continuous. If $\{(\sigma_n, r_n)\}$ is a sequence of points in Ξ converging to $(\sigma, 0)$ and $y \in Y$, the intermediate value theorem implies that for each n there is a $r'_n \in (0, r_n)$ such that $\psi(\sigma_n, r_n, y) = \omega^* + \frac{\partial \kappa}{\partial t}(\sigma_n, r'_n, y)$, so the continuity of $\frac{\partial \kappa}{\partial r}$ implies that $\psi(\sigma_n, t_n, y) \to \psi(\sigma, 0, y)$. Thus ψ is continuous.

We define a function $B_p : \overline{\Xi} \to \Delta(H)$ by letting $B_p(\sigma, r)$ be the element of $\Delta(H)$ that assigns probability

$$\sum_\theta (\omega^*_\theta + r\sigma_\theta) q_\theta(y|p(\omega^* + r\sigma))$$

to each $\psi(\sigma, r, y)$. Clearly B_p is continuous.

Our first main result is as follows. The proof is given at the end of Section 3, after the supporting statistical result has been developed.

Theorem 1. *If, for all $\sigma \in S$, $\int_H \ln \|\omega - \omega^*\| \, dB_p(\sigma, 0) < 0$, then for any sufficiently small neighborhood U of ω^*, if the probability that $\tilde{\omega}_0 \in U$ is positive, then there is a positive probability that $\tilde{\omega}_t \to \omega^*$.*

Turning to the second question, we now consider the problem of maximizing the expectation of

$$\sum_{t=0}^{\infty} \delta^t u(\tilde{\omega}_t, \tilde{a}_t)$$

where $u : \Omega \times A \to \mathbb{R}$ is a continuous function. In many applications $u(\omega, a)$ will be the expectation of a reward function $R : A \times Y \to \mathbb{R}$:

$$u(\omega, a) = \sum_\theta \omega_\theta \sum_y R(a, y) q_\theta(y|a).$$

Our second main result, which is stated and proved at the end of Section 5, has the following intuition: under natural and easily verified conditions, the optimal policy p_δ, and its derivative, will vary continuous as δ varies in a neighborhood of zero. Provided that the dimension of Ω is the same as the dimension of A, it follows from the theory of the degree that for small $\delta > 0$, some point near ω^* is mapped to a^*. Since choosing a^* minimizes learning, it cannot be optimal to do so for positive δ unless it is also myopically optimal, so we have $p_\delta(\omega^*) = a^*$ for small positive δ. Since the derivative of p varies continuously with δ, if $\int_H \ln \|\omega - \omega^*\| \, dB_{p_0}(\sigma, 0) < 0$ for all $\sigma \in S$, then for small positive δ it is also the case that $\int_H \ln \|\omega - \omega^*\| \, dB_{p_\delta}(\sigma, 0) < 0$ for all $\sigma \in S$.

3 Bounds on Escape Probabilities

For the analysis in the section, prior to the proof of Theorem 1, $(\Omega, \mathcal{F})$ may be any measurable space. Let $\Delta(\Omega)$ be the set of probability measures on Ω. We study a stationary Markov process $\tilde{\omega}_0, \tilde{\omega}_1, \tilde{\omega}_2, \ldots$ in Ω with Markov kernel $P : \Omega \to \Delta(\Omega)$. That is, for all $E \in \mathcal{F}$, $P(E|\cdot) : \Omega \to [0,1]$ is measurable, and for all $\omega_0, \ldots, \omega_t$ we have

$$\Pr(\tilde{\omega}_{t+1} \in E | \tilde{\omega}_0 = \omega_0, \ldots, \tilde{\omega}_t = \omega_t) = P(E|\omega_t).$$

(More formally, $(\omega_1, \ldots, \omega_t) \mapsto P(\cdot|\omega_t)$ is a version of conditional probability for the distribution of of $(\tilde{\omega}_1, \ldots, \tilde{\omega}_t, \tilde{\omega}_{t+1})$.) Let $\ell : \Omega \to \mathbb{R}$ be a measurable functions. We study conditions on P and ℓ that imply that there is a positive probability that the sequence $\ell(\tilde{\omega}_0), \ell(\tilde{\omega}_1), \ell(\tilde{\omega}_2), \ldots$ never gets above zero, in which case $\ell(\tilde{\omega}_t) \to -\infty$ almost surely.

We begin with a technical result:

Lemma 1. *Let $\tilde{x}$ be a random variable with cumulative distribution function Φ. If $\mathbf{E}(e^{\gamma \tilde{x}}) < \infty$ for some $\gamma > 0$ and $\mathbf{E}(\tilde{x}) < 0$, then there exist $C, \overline{\beta} > 0$ such that*

$$1 - \Phi(-y) < C e^{\beta y} (1 - \int_{-\infty}^{-y} e^{\beta x} \, \Phi(dx))$$

for all all $\beta \in (0, \overline{\beta})$ and $y \leq 0$.

Proof. One can easily show that for any $M > 0$,

$$\frac{\mathbf{E}(e^{\gamma \tilde{x}} | -M \leq \tilde{x} \leq M)}{\gamma} \to e^{\mathbf{E}(\tilde{x} | -M \leq \tilde{x} \leq M)}$$

as $\gamma \to 0$, and from this it follows easily that there is a $\overline{\beta} > 0$ such that $\mathbf{E}(e^{\beta \tilde{x}}) < 1$ for all all $\beta \in (0, \overline{\beta})$. We can now choose

$$C = \sup_{0 \leq \beta \leq \overline{\beta}, -\infty < y \leq 0} \frac{(1 - \Phi(-y)) e^{-\beta y}}{1 - \int_{-\infty}^{-y} e^{\beta x} \, \Phi(dx)}.$$

This supremum is not infinite because $(1 - \Phi(-y)) e^{-\beta y} \to 0$ as $y \to -\infty$, since otherwise $\mathbf{E}(e^{\beta \tilde{x}}) = \infty$. □

Proposition 1. *Suppose that $\Phi_1, \ldots, \Phi_K : \mathbb{R} \to [0,1]$ are cumulative distribution functions such that:*

(a) For each k, if $\tilde{x}_k$ is distributed according to Φ_k, then $\mathbf{E}(\tilde{x}_k) < 0$.

(b) for each $\omega \in \Omega$ such that $\ell(\omega) < 0$ there is some k such that $P(\{\omega' \in \Omega : \ell(\omega') \geq \ell(\omega) + x\} | \omega) \leq 1 - \Phi_k(x)$ for all $x \in \mathbb{R}$.

Then there are $C, \beta > 0$ such that for all ω_0 with $\ell(\omega_0) < 0$ we have

$$\Pr(\ell(\tilde{\omega}_t) \to -\infty | \tilde{\omega}_0 = \omega_0) = \Pr(\ell(\tilde{\omega}_t) < 0 \textit{ for all sufficiently large } t | \tilde{\omega}_0 = \omega_0)$$

$$\geq \Pr(\ell(\tilde{\omega}_t) < 0 \textit{ for all } t | \tilde{\omega}_0 = \omega_0) > 1 - Ce^{\beta \ell(\omega_0)}.$$

Proof. In view of Lemma 1, there are $C, \beta > 0$ such that

$$1 - \Phi(-y) < Ce^{\beta y}\Big(1 - \int_{-\infty}^{-y} e^{\beta x}\, \Phi_k(dx)\Big) \qquad (*)$$

for all all k and $y \leq 0$. For each $T = 0, 1, 2, \ldots$ let $p_T : \Omega \to [0, 1]$ be the function

$$p_T(\omega_0) = \Pr\big(\ell(\tilde{\omega}_T) \geq 0 \text{ for some } t = 0, \ldots, T \,|\, \tilde{\omega}_0 = \omega_0\big).$$

It suffices to show that for a given ω_0 such that $\ell(\omega_0) < 0$ we have $p_T(\omega_0) \leq Ce^{\beta \ell(\omega_0)}$ for all T. This is obviously true when $T = 0$, so, by induction, we may suppose that it has already been established with $T - 1$ in place of T. As per (b), choose k such that $P(\{\omega' \in \Omega : \ell(\omega') \geq \ell(\omega) + x\}|\omega) \leq 1 - \Phi_k(x)$ for all $x \in \mathbb{R}$. Then

$$p_T(\omega_0) = P(\{\omega : \ell(w) \geq 0\}|\omega_0) + \int_{\{\omega : \ell(\omega) < 0\}} p_{T-1}(\omega)\, P(d\omega|\omega_0)$$

$$\leq 1 - \Phi_k(-\ell(\omega_0)) + \int_{-\infty}^{-\ell(\omega_0)} Ce^{\beta(\ell(\omega_0)+x)}\, \Phi_k(dx)$$

$$= 1 - \Phi_k(-\ell(\omega_0)) + Ce^{\beta \ell(\omega_0)}\Big(\int_{-\infty}^{-\ell(\omega_0)} e^{\beta x}\, \Phi_k(dx) - 1\Big) + Ce^{\beta \ell(\omega_0)}.$$

Now $(*)$ implies that $p_T(\omega_0) \leq Ce^{\beta \ell(\omega_0)}$. □

Proof of Theorem 1. For each $\sigma \in S$, if ω is distributed according to $B_p(\sigma, 0)$, then the expectation of $\ln \|\omega - \omega^*\|$ is negative. Since $B_p(\sigma, 0)$ has finite support, we can define a distribution on $\mathbb{R}$ by assigning the same probabilities to numbers slightly larger than the $\ln \|\omega - \omega^*\|$, so that the mean of this distribution is negative. Then for any (σ', r') in some neighborhood of $(\sigma, 0)$ in $\overline{\Xi}$, this distribution also first order stochastically dominates the distribution of $\ln \|\omega - \omega^*\|$ when ω is distributed according to $B_p(\sigma', r')$. Since S is compact, it follows that there is a finite collection of neighborhoods that covers $S \times \{0\}$. The union of these neighborhoods is a neighborhood of $S \times \{0\}$, so it contains $S \times [0, \overline{r}]$ for some $\overline{r} > 0$. Let $\ell : \Omega \setminus \{\omega^*\} \to \mathbb{R}$ be the function

$$\ell(\omega) = \ln \|\omega - \omega^*\| - \ln \overline{r}.$$

At this point we have verified the hypotheses of Proposition 1, and it implies the desired conclusion. □

4 Manifolds with Corners

We will apply methods from differential topology, but the most important example for our purposes, namely the simplex, is not a manifold with boundary. It is a manifold with corners, which is a slightly more general concept that is much less popular in the mathematical literature. This section describes the relevant concepts, which are standard (cf. Hirsch (1976)) in this setting.

Fix a degree of differentiability $1 \leq r \leq \infty$. We first recall that an m-dimensional C^r *manifold* is a topological space M together with a collection $\{\varphi_g : U_g \to \mathbb{R}^m_{\geq}\}_{g \in G}$ where $\{U_g\}$ is an open cover of M, each φ_g is a homeomorphism between U_g and $\varphi_g(U_g)$, each $\varphi_g(U_g)$ is an open subset of $\mathbb{R}^m$, and all the maps $\varphi_g \circ \varphi_{g'}^{-1}$ are C^r on their domains of definition. The collection $\{\varphi_g : U_g \to \mathbb{R}^m_{\geq}\}_{g \in G}$ is said to be a C^r *atlas* for M.

Recall that for any $D \subset \mathbb{R}^m$, a function $f : D \to \mathbb{R}$ is differentiable if there is a differentiable extension of f to an open superset of D. We say that D is a *differentiation domain* if, for any differentiable $f : D \to \mathbb{R}$ and any two differentiable extensions $f' : U' \to \mathbb{R}$ and $f'' : U'' \to \mathbb{R}$, the derivatives of f' and f'' agree at all points of D. For example, the positive orthant $\mathbb{R}^m_{\geq}$ is a differentiation domain.

An m-dimensional C^r *manifold with corners* is a topological space M with a collection $\{\varphi_g : U_g \to \mathbb{R}^m_{\geq}\}_{g \in G}$ where now each $\varphi_g(U_g)$ is an open subset of $\mathbb{R}^m_{\geq}$, and, as above, $\{U_g\}$ is an open cover of M, each φ_g is a homeomorphism between U_g and $\varphi_g(U_g)$, and all the maps $\varphi_g \circ \varphi_{g'}^{-1}$ are C^r on their domains of definition. The collection $\{\varphi_g : U_g \to \mathbb{R}^m_{\geq}\}_{g \in G}$ is said to be a C^r *atlas* for M. A set $D \subset M$ is a *differentiation domain* if, for each $g \in G$, $\varphi_g(D \cap U_g)$ is a differentiation domain.

The simplex provides a simple concrete example: let $\Omega = \{\, (x_0, \ldots, x_m) \in \mathbb{R}^{m+1} : \sum_g x_g = 1 \,\}$. A C^∞ atlas for Ω is given by letting $U_g = \{\, x \in \Omega : x_g > 0 \,\}$ for each $g = 0, \ldots, m$, and letting $\varphi_g : U_g \to \mathbb{R}^m_{\geq}$ be the map

$$\varphi(x) = (x_0, \ldots, x_{g-1}, x_{g+1}, \ldots, x_m).$$

Suppose that $D \subset M$ is a differentiation domain. If $0 \leq s \leq r$, a function $f : M \to \mathbb{R}$ is said to be C^s if each map $f \circ \varphi_g^{-1} : \varphi_g(D \cap U_g) \to \mathbb{R}$ is C^s. Let $C^s(D)$ be the space of such maps.

Now suppose that D is compact. We wish to define suitable metrics on the spaces $C^s(D)$. Suppose that $K_1, \ldots, K_L$ is a collection of compact subsets of M whose interiors cover D, and for each $\ell = 1, \ldots, L$ there is a $g_\ell \in \{1, \ldots, G\}$ such that $K_\ell \subset U_{g_\ell}$. For $f, f' \in C^s(D)$ let

$$d_s^D(f, f') = \max \left| \frac{\partial^s (f \circ \varphi_{g_\ell}^{-1})}{\partial x_{i_1} \cdots \partial x_{i_t}} (\varphi_{g_\ell}(p)) - \frac{\partial^s (f' \circ \varphi_{g_\ell}^{-1})}{\partial x_{i_1} \cdots \partial x_{i_t}} (\varphi_{g_\ell}(p)) \right|$$

where the maximum is over all $\ell = 1, \ldots, L$, $p \in K_\ell$, $t = 0, \ldots, s$, and $i_1, \ldots, i_t = 1, \ldots, m$. It is easy to verify that d_s^D is a metric.

Note that d_r^D is the metric derived from the norm

$$\|f\| = \max_{\ell,p,s,i_1,\ldots,i_s,j} \left| \frac{\partial^s(\psi_{h_\ell j} \circ f \circ \varphi_{g_\ell}^{-1})}{\partial x_{i_1} \cdots \partial x_{i_s}}(\varphi_{g_\ell}(p)) \right|.$$

Consequently d_r^D respects multiplication by scalars in the sense that $d^D(\alpha f, \alpha f') = |\alpha| \cdot d^D(f, f')$ for all $\alpha \in \mathbb{R}$. It is well known that with this norm $C^r(D)$ is complete and thus a Banach space.

In addition to defining topologies, these metrics provide a notion of what it means for a function between these spaces to be locally Lipschitz. For general metric spaces (X, d) and (Y, e) a function $f : X \to Y$ is *Lipschitz at* $x \in X$ if there is a neighborhood U of x and a $\Lambda > 0$ such that $e(f(x'), f(x'')) \leq \Lambda d(x', x'')$ for all $x', x'' \in U$, and f is *locally Lipschitz* if it is Lipschitz at each point in X. Of course the metric d_s^D depends on the atlases, the sets K_ℓ, and the integers g_l, and whether a function to or from $C^s(D)$ is locally Lipschitz must not depend on this data. A composition of two locally Lipschitz functions is locally Lipschitz, so this follows from the fact that if d_s^D and $\tilde{d}_s^D$ are two such metrics, then the identity function is locally Lipschitz when the domain has the metric $d_s^{(M,N)}$ and the range has metric $\tilde{d}_s^{(M,N)}$. The ideas underlying the proof of this (intersections of compact sets are compact, the chain rule, continuous functions with compact domains are bounded) are elementary, but a detailed description of the argument would be rather cumbersome, so we leave the verification to the reader.

Let $C^s(D, \mathbb{R}^n)$ denote the n-fold cartesian product of $C^s(D)$, and let the metric $d_s^{(D,\mathbb{R}^n)}$ be defined by

$$d_s^{(D,\mathbb{R}^n)}(f, f') = d_s^D(f_1, f_1') + \cdots + d_s^D(f_n, f_n').$$

If $U \subset \mathbb{R}^n$ is open, let $C^s(D, U)$ be the set of $f \in C^s(D, \mathbb{R}^n)$ with $f(D) \subset U$. Usually we write $C(D)$ in place of $C^0(D)$ and $C(D, U)$ in place of $C^0(D, U)$.

5 Perturbation Methods

In this section we analyze a dynamic programming problem, aiming at results stating that the optimal policy function and the value function vary continuously, in the topologies described in the last section, as the discount factor δ varies in a neighborhood of $\delta = 0$.

Let the space of *states* Ω be a compact m-dimensional C^2 manifold with corners, and let the space of actions A be an open subset of $\mathbb{R}^n$. Then A is a C^2 manifold, hence a C^2 manifold with corners, and the cartesian product of two C^2 manifolds with corners

is easily seen to be another C^2 manifold with corners, so $\Omega \times A$ is a C^2 manifold with corners. Let $D \subset \Omega \times A$ be a compact differentiation domain. For each $\omega \in \Omega$ let $D(\omega)$ be the set of a such that $(\omega, a) \in D$. In what follows we will be concerned only with maximization over the sets $D(\omega)$. In applications the relevant notion of maximization will often extend over all of A, and the particular context will determine the considerations that insure that the local maximizers analyzed below are in fact global optima.

Let $u_0 : D \to \mathbb{R}$ be a C^2 function whose second order derivatives are Lipschitz functions. We assume that for each $\omega \in \Omega$ there is a unique maximizer $p_{u_0}(\omega)$ of $u_0(\omega, \cdot) : D(\omega) \to \mathbb{R}$ such that (ω, a) is in the interior of D and the second order necessary conditions for maximization hold strictly. We begin the analysis by studying how the optimal policy and the value vary when we perturb u_0 in $C^2(D)$.

There is an operator J that maps $C(D)$ to real valued functions on Ω that is defined by setting

$$J(u')(\omega) = \max_{a \in D(\omega)} u'(\omega, a).$$

Proposition 2. *There is a neighborhood $W \subset C^2(D)$ of u such that:*

(a) For each $u' \in W$ and each ω there is a unique maximizer $p_{u'}(\omega)$ of $u'(\omega, \cdot)$ that is in the interior of A, at which the second order necessary conditions for maximization hold strictly.

(b) For each $u' \in W$, $p_{u'} : \Omega \to A$ is C^1.

(c) The operator $u' \mapsto p_{u'}$ is a locally Lipschitz function from W to $C^1(\Omega, A)$.

(d) $J(W) \subset C^2(\Omega)$.

(e) $J|_W$ is locally Lipschitz.

Proof. Consider a particular $\omega \in \Omega$. For any neighborhood of $p_u(\omega)$, if u' is sufficiently close to u in the metric d_0^D, then all the maximizers of $u'(\omega, \cdot)$ will lie in that neighborhood. By choosing this neighborhood appropriately, one can insure that if u' is sufficiently close to u in the metric d_2^D, then there will be a unique point in the neighborhood at which the first order conditions are satisfied, with the second order conditions holding strictly. Thus, if u' is sufficiently close to u in the metric d_2^D, then (a) holds, and (b) follows from the implicit function theorem.

Using the fact that the second order conditions hold strictly, it is easy to see that the map $u' \mapsto p_{u'}$ is Lipschitz when u' is sufficiently close to u in the metric d_2^D and the range has the metric $d_0^{(\Omega, A)}$.

In order to simplify the notation, for the rest of the proof we assume that $\Omega = A = [0,1]$. It will be clear that all the steps in the argument generalize in a straightforward manner. Fully differentiating the equation

$$\frac{\partial u'}{\partial a}(\omega, p_{u'}(\omega)) = 0$$

and rearranging gives

$$\frac{dp_{u'}}{d\omega}(\omega) = -\frac{\frac{\partial^2 u'}{\partial \omega \partial a}(\omega, p_{u'}(\omega))}{\frac{\partial^2 u'}{\partial a^2}(\omega, p_{u'}(\omega))}.$$

We can now decompose the difference between $\frac{dp_u}{d\omega}$ and $\frac{dp_{u'}}{d\omega}$ into two parts: i) the consequence of replacing $p_u(\omega)$ with $p_{u'}(\omega)$ in the expression above; ii) the consequence of replacing the various second partial derivatives of u with the corresponding second partials of u'. As for i), it is bounded by a multiple of $d_2^D(u', u)$ because $d_0^{(\Omega,A)}(p_{u'}, p_u)$ is bounded by a multiple of this distance (as we noted above) and the second partials of u are Lipschitz by assumption. Of course ii) is bounded by a constant multiple of $d_2^D(u', u)$ because this distance bounds the differences in the relevant second partials, and the expression above is a locally Lipschitz function of these partials. Thus (c) holds.

We have

$$V_{u'}(\omega) = u'(\omega, p_{u'}(\omega)),$$

$$V'_{u'}(\omega) = \frac{\partial u'}{\partial \omega}(\omega, p_{u'}(\omega)),$$

$$V''_{u'}(\omega) = \frac{\frac{\partial^2 u'}{\partial \omega^2}(\omega, p_{u'}(\omega)) \cdot \frac{\partial^2 u'}{\partial a^2}(\omega, p_{u'}(\omega)) - \frac{\partial^2 u'}{\partial \omega \partial a}(\omega, p_{u'}(\omega))^2}{\frac{\partial^2 u'}{\partial a^2}(\omega, p_{u'}(\omega))},$$

by virtue of, respectively, the definition of $V_{u'}$, the envelope theorem, and total differentiation of the second equation followed by substituting the formula for $\frac{dp_{u'}}{d\omega}$ above. Evidently $V_{u'}$ is C^2, so (d) holds. In addition, an argumnent similar to the one given above decomposing the differences between V'_u and $V'_{u'}$ and between V''_u and $V''_{u'}$ into the effects of replacing p_u with $p_{u'}$ and the effects of replacing u with u' establishes (e). □

Let $q : \Omega \times A \to \Delta(\Omega)$ be a given continuous function. The dynamic program is to maximize the expectation of

$$\sum_{t=0}^{\infty} \delta^t u(\tilde{\omega}_t, \tilde{a}_t)$$

where $\tilde{\omega}_0 = \omega_0$ almost surely, $\tilde{\omega}_t$ is known at the time $\tilde{a}_t$ is chosen, and, conditional on $\tilde{\omega}_t$ and $\tilde{a}_t$, $\tilde{\omega}_{t+1}$ has the distribution $q(\tilde{\omega}_t, \tilde{a}_t)$. Let $V_0(\omega_0) = u(\omega_0, p_u)$ be the value of this problem when $\delta = 0$.

We analyze this using the standard methodology of dynamic programming, by analyzing the value of the problem as a function of ω_0. There is an operator

$$K : C(\Omega) \to C(\Omega \times A) \quad \text{given by} \quad K(V)(\omega, a) = \int_\Omega V(\cdot)\, dq(\omega, a).$$

Proposition 3. *Assume there a neighborhood $Z \subset C^2(\Omega)$ of V_u such that $K(Z) \subset C^2(\Omega \times A)$ and $K|_Z$ is Lipschitz. Then there is a $\overline{\delta} > 0$ such that if $0 \leq \delta < \overline{\delta}$, then the optimal policy $p_\delta : \Omega \to A$ is C^1. In addition, the map $\delta \mapsto p_\delta$ is Lipschitz relative to the metric $d_1^{(\Omega,A)}$.*

Proof. After replacing W from Proposition 2 with a smaller neighborhood of u, we may assume that $J(W) \subset Z$. For $\delta \in \mathbb{R}$ let $L_\delta : W \to C^2(\Omega)$ be the operator

$$L_\delta(u') = u + \delta \cdot K(J(u')).$$

Since d^Ω and d^D respect multiplication by scalars, there is a $\overline{\delta} > 0$ and a constant $\lambda \in (0, 1)$ such that for all $\delta \in (0, \overline{\delta})$, L_δ is a contraction with modulus of contraction at most λ. Since $C^2(\Omega \times A)$ is a Banach space, and consequently complete, in this circumstance L_δ must have a fixed point u_δ. Let p_δ be the optimal policy function for the discounted problem with discount factor δ. A standard argument based on iteratively applying the operator $L_{\delta'}$ to u_δ shows that the map $\delta \mapsto u_\delta$ is Lipschitz with Lipschitz constant $1/(1 - \lambda)$. In view of the last result, it follows that the map $\delta \to p_\delta$ is also Lipschitz. □

We can now state and prove the second main result.

Theorem 2. *Suppose that for each $\omega \in \Omega$, $p_0(\omega)$ is the unique maximizer of $u(\omega, \cdot)$, and that the second order conditions for maximization hold strictly there. Assume that $\int_H \ln \|\omega - \omega^*\|\, dB_{p_0}(\sigma, 0) < 0$ for all $\sigma \in S$. Assume also that there is a compact neighborhood $D \subset \Omega \times A$ of the graph of p_0 that is a differentiation domain and an $\varepsilon > 0$ such that $u(\omega, a) < u(\omega, p_0(\omega)) - \varepsilon$ for all $(\omega, a) \in (\Omega \times A) \setminus D$. Assume there a neighborhood $Z \subset C^2(\Omega)$ of V_u such that $K(Z) \subset C^2(\Omega \times A)$ and $K|_Z$ is Lipschitz. Finally, assume that the dimension of Ω is the same as the dimension of A, and that the derivative of p_0 at ω^* is nonsingular. Then there is a $\overline{\delta} > 0$ such that for all $\delta \in [0, \overline{\delta})$, the optimal policy p_δ is C^1, $p_\delta(\omega^*) = a^*$, and $\int_H \ln \|\omega - \omega^*\|\, dB_{p_\delta}(\sigma, 0) < 0$ for all $\sigma \in S$.*

Proof. If we restricted the decision maker to policies with graphs lying in D, all the results above would be available. But in fact it is not hard to show that the assumption that actions outside of $D(\omega)$ are myopically ε-suboptimal implies, for sufficiently small $\delta > 0$, that the optimal policies with graphs lying in D are, in fact optimal.

Therefore $\delta \to p_\delta$ is continuous when the range has the C^1 topology. By the theory of the degree, continuity relative to the C^0 topology implies that for sufficiently small δ there will be some point near ω^* that is mapped to a^*. Since a^* minimizes learning, it cannot be optimal for small positive δ unless it is also myopically optimal, and since the derivative of p_0 at ω^* is nonsingular, it follows that we must have $p_\delta(\omega^*) = a^*$ for small $\delta > 0$.

Since p_δ varies continuously with δ in the C^1 sense, $B_{p_\delta}(\sigma, 0)$ is jointly continuous as a function of δ and σ, and from this it follows that for small positive δ we have $\int_H \ln \|\omega - \omega^*\| \, dB_{p_\delta}(\sigma, 0) < 0$ for all $\sigma \in S$. □

6 Applying the Results

In this section we explain how $B_p(\cdot, 0)$ is computed. This computation is required by any concrete application of the result. In addition, we are able to analyze the extent to which increasing the amount of experimentation affects the probability of falling into the learning trap. When Ω and A are 1-dimensional, it is easy to see that learning traps are in fact possible, but when the common dimension is greater than one it is unclear whether it can actually happen that $\int_H \ln \|\omega - \omega^*\| \, dB_{p_0}(\sigma, 0) < 0$ for all $\sigma \in S$. We analyze an example which provides an inconclusive result.

Let $q(y|a^*)$ denote the common value of $q_\theta(y|a^*)$, which is the probability of observing y when the uninformative action a^* is chosen. For $\sigma \in S$ and $y \in Y$ let

$$\rho_\theta(\sigma, y) = \frac{1}{q(y|a^*)} \cdot \frac{\partial q_\theta(y|p(\omega^* + r\sigma))}{\partial r}\Big|_{r=0}.$$

Using elementary calculus, it is not hard to show that

$$\psi(\sigma, 0, y) = \omega^* + \sigma + \nu(\sigma, y)$$

where the components of the vector $\nu(\sigma, y)$ are

$$\nu_\theta(\sigma, y) = \rho_\theta(\sigma, y) - \omega^*_\theta \sum_{\theta' \in \Theta} \rho_{\theta'}(\sigma, y).$$

Then $B_p(\sigma, 0)$ assigns probability $q(y|a^*)$ to each $\psi(\sigma, 0, y)$.

A common intuition is that more aggressive experimentation will result in fewer learning traps and a lower probability of falling into one. One way to think about this is to consider replacing p with another policy p_α mapping ω^* to a^* whose derivative at ω^* is the derivative of p multiplied by the scalar $\alpha > 0$. This replacement results in the numbers $\rho_\theta(\sigma, y)$ and the vector $\nu(\sigma, y)$ being multiplied by the same scalar. For large

values of α we have $\|\sigma + \alpha\nu(\sigma, y)\| > \|\sigma\|$ for all y, in which case ω^* is "repelling" rather than "attracting."

It is also interesting to consider what happens when α is small. Consider two vectors v and w in the hyperplane through the origin parallel to H with $\|v\| = 1$. If $g(s) = \ln\|v + sw\|$, then, by elementary calculus,

$$g'(s) = \frac{\langle v + sw, w\rangle}{\|v + sw\|^2}$$

and

$$g''(s) = \frac{\|v + sw\|^2 \cdot \|w\|^2 - 2\langle v + sw, w\rangle^2}{\|v + sw\|^4},$$

so that $g'(0) = \langle v, w\rangle$ and $g''(0) = \|w\|^2 - 2\langle v, w\rangle^2$. In view of these results we have

$$\frac{d\big(\int_H \ln\|\omega - \omega^*\|\, dB_{p_\alpha}(\sigma, 0)\big)}{d\alpha}\Big|_{\alpha=0} = \sum_y q(y|a^*)\langle\sigma, \nu(\sigma, y)\rangle = 0$$

and

$$\frac{d^2\big(\int_H \ln\|\omega - \omega^*\|\, dB_{p_\alpha}(\sigma, 0)\big)}{d\alpha^2}\Big|_{\alpha=0} = \sum_y q(y|a^*)\big(\|\nu(\sigma, y)\|^2 - 2\langle\sigma, \nu(\sigma, y)\rangle^2\big). \qquad (*)$$

This finding suggests that when there is already a small amount of experimentation as one moves away from ω^*, the effect of further reducing experimentation, by replacing p with p_α where $\alpha < 1$, is primarily to slow the process down, without changing the probability of eventual convergence to ω^*. If we think of the process as akin to a Brownian motion, the result of the replacement is to multiply the drift and the instantaneous variance by α^2; as Callander (2011) points out in a setting with Brownian motion, the result of such a replacement is a rescaling of the process that does not change its limiting properties.

The last result is also interesting from a different point of view. When Ω and A are 1-dimensional, the right hand side of $(*)$ reduces to $-\sum_y q(y|a^*)\|\nu(\sigma, y)\|^2$ because $\nu(\sigma, y)$ is necessarily a scalar multiple of σ. Thus, in this case, there is a positive probability of convergence to ω^* when α is sufficiently small. When $n \geq 2$ this is no longer the case, and in fact a very interesting question is for which values of n can it be the case that the right hand side of $(*)$ is negative for all $\sigma \in S$.

We now analyze a concrete example. Let $\Theta = \{\theta_1, \theta_2, \theta_3\}$, let A be the interior of Ω, let p be the identity function, and let $\omega^* = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Let $Y = \{y_1, y_2, y_3\}$, and let

$$q_{\theta_i}(y_j|a) = \begin{cases} \frac{1}{3} + \frac{2}{3}(a_i - \frac{1}{3}), & i = j, \\ \frac{1}{3} - \frac{1}{3}(a_i - \frac{1}{3}), & i \neq j. \end{cases}$$

Then

$$\frac{\partial q_\theta(y|p(\omega^*+r\sigma))}{\partial r}\Big|_{r=0} = \begin{cases} \frac{2}{3}\sigma_i, & i=j, \\ -\frac{1}{3}\sigma_i, & i\neq j. \end{cases}$$

We compute that

$$\sum_i \rho_{\theta_i}(\sigma, y_j) = \sigma_j$$

because $\sigma_1+\sigma_2+\sigma_3=0$. Therefore

$$\nu_{\theta_i}(\sigma, y_j) = \begin{cases} \sigma_i, & i=j, \\ -\sigma_i-\sigma_j, & i\neq j. \end{cases}$$

Since $-\sigma_1-\sigma_2=\sigma_3$, and similarly in the other cases, we have

$$\nu(\sigma,y_1)=(\sigma_1,\sigma_3,\sigma_2), \quad \nu(\sigma,y_2)=(\sigma_3,\sigma_2,\sigma_1), \quad \nu(\sigma,y_3)=(\sigma_2,\sigma_1,\sigma_3).$$

Now we have

$$\sum_y q(y|a^*)\big(\|\nu(\sigma,y)\|^2-2\langle\sigma,\nu(\sigma,y)\rangle^2\big) = 1-\tfrac{2}{3}\big((\sigma_1^2+2\sigma_2\sigma_3)^2+(\sigma_2^2+2\sigma_1\sigma_3)^2+(\sigma_3^2+2\sigma_1\sigma_2)^2\big)$$

$$= 1-\tfrac{2}{3}\Big(\sum_i \sigma_i^4 + 4\sum_i \sigma_i(\sigma_1+\sigma_2+\sigma_3) + 4\sum_{j\neq k}\sigma_j^2\sigma_k^2\Big)$$

$$= 1-\tfrac{2}{3}\Big(\sum_i \sigma_i^4 + 4\sum_{j\neq k}\sigma_j^2\sigma_k^2\Big) = 1-\tfrac{2}{3}\big((\sigma_1^2+\sigma_2^2+\sigma_3^2)^2 + 2\sum_{j\neq k}\sigma_j^2\sigma_k^2\big)$$

$$= \tfrac{1}{3} - \tfrac{4}{3}\sum_{j\neq k}\sigma_j^2\sigma_k^2 = -\tfrac{1}{3}+\tfrac{2}{3}(\sigma_1^4+\sigma_2^4+\sigma_3^4).$$

(The last equality follows from $\sigma_1^2\sigma_2^2+\sigma_1^2\sigma_3^2=\sigma_1^2(-\sigma_1^2)$ and symmetric equations.) Some manipulation of the equations

$$(\sigma_1^2+\sigma_2^2+\sigma_3^2)^2=1 \quad \text{and} \quad (\sigma_1+\sigma_2+\sigma_3)^4=0$$

shows that $\sigma_1^4+\sigma_2^4+\sigma_3^4=\frac{1}{2}$ is a consequence of these equations. Thus

$$\frac{d^2\big(\int_H \ln\|\omega-\omega^*\|\, dB_{p_\alpha}(\sigma,0)\big)}{d\alpha^2}\Big|_{\alpha=0} = 0$$

for all $\sigma\in S$. One could attempt to push the analysis further by looking at the third and fourth derivatives, or one could turn to other examples. With respect to this issue, this draft is indeed "preliminary and incomplete."

7 Concluding Remarks

We have provided an analysis of Bayesian learning traps that gives sufficient conditions for them to occur, and we have shown that if myopically optimal policies allow them, then so do the optimal policies of decision makers with small positive discount factors. A major unresolved issue is whether these conditions can actually occur in multidimensional settings.

Our analysis in this paper is restricted in several ways.

A natural direction of generalization is to situations in which the dimension of the set of uninformative actions is a submanifold of A of positive dimension, and the dimension of the space of beliefs may be greater than the codimension of the set of uninformative actions. When these objects and the policy function are "well behaved," the set of beliefs mapped to uninformative actions will be a submanifold of Ω, and the question becomes whether there can be a positive probability that the sequence of beliefs converges to a point in this submanifold. One can anticipate certain additional technical complications, but at this point there seems to be little reason to expect the qualitative properties of the results to change.

A major direction for generalization is to consider the possibility that Y is infinite. In particular, the case of normally distributed shocks is a central concern. Again, significant additional complications can be foreseen, but at this point we are not aware of any insuperable obstacles.

Finally, an economically important possibility is that learning might be incomplete because there is a positive probability of convergence to a belief whose support is not all of Θ. As with the other extensions described above, this appears to present certain challenges, which most likely can be overcome.

References

Aghion, P., Bolton, P., Harris, C., and Julien, B. (1991). Optimal learning by experimentation. *Review of Economic Studies*, 58:621–654.

Baker, S. and Mezzetti, C. (2011). A theory of rational jurisprudence. Mimeo, Washington University in St.Louis and University of Melbourne.

Banerjee, A. V. (1992). A simple model of herd behavior. *Quarterly Journal of Economics*, 107:73–88.

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 gittens index. *Econometrica*, 62:687–694.

Berentsen, A., Bruegger, E., and Loertscher, S. (2008). Learning, public goods provision and the information trap. *Journal of Public Economics*, 92:998–1010.

Berry, D. and Fristedt, B. (1985). *Bandit Problems: Sequential Allocation of Experiments*. Monographs on Statistics and Applied Probability. Chapman and Hall, London.

Bikhchandani, S., Hirshleifer, D., and Welch, I. (1992). A theory of fads, fashion, custom, and cultural change in informational cascades. *Journal of Political Economy*, 100:992–1026.

Callander, S. (2011). Searching and learning by trial and error. *American Economic Review*, 101:2277–2308.

Easley, D. and Kiefer, N. (1988). Controlling a stochastic process with unknown parameters. *Econometrica*, 56:1045–1064.

Ellison, G. and Fudenberg, D. (1995). Word-of-moth communication and social learning. *Quarterly Journal of Economics*, 110:93–126.

Gittens, J. and Jones, D. (1974). A dynamic allocation index for the sequential allocation of experiments. In Gani, J., editor, *Progress in Statistics*, pages 241–266. North Holland, Amsterdam.

Harrison, J.-M., Keskina, N.-B., and Zeevi, A. (2011). Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. forthcoming in *Management Science*.

Hirsch, M. W. (1976). *Differential topology*. Springer-Verlag, New York. Graduate Texts in Mathematics, No. 33.

Kihlstrom, R. E., Mirman, L. J., and Postlewaite, A. (1984). Experimental consumption and the 'Rothschild effect'. In Boyer, M. and Kihlstrom, R. E., editors, *Bayesian Models in Economic Theory*. Elsevier, Amsterdam.

Laslier, J.-F., Trannoy, A., and van der Straeten, K. (2003). Voting under ignorance of job skills of unemployed: the overtaxation bias. *Journal of Public Economics*, 87:595–626.

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

Piketty, T. (1995). Social mobility and redistributive politics. *Quarterly Journal of Economics*, 110:551–584.

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

Rothschild, M. and Stiglitz, J. (1984). A nonconcavity in the value of information. In Boyer, M. and Kihlstrom, R., editors, *Bayesian Models of Economic Theory*, pages 33–52. Elsevier, Amsterdam.

Smith, L. and Sørensen, P. (2000). Pathological outcomes of observational learning. *Econometrica*, 68:371–398.

Smith, L. and Sørensen, P. (2005). Informational herding and optimal experimentation. Working Paper 05-13, University of Copenhagen.

Smith, L. and Sørensen, P. (2011). Observational learning. In Durlauf, S. and Blume, L., editors, *The New Palgrave Dictionary of Economics Online Edition*, pages 29–52. Palgrave Macmillan, New York.