# Some People Never Learn, Rationally

Simon Loertscher
University of Melbourne*

Andrew McLennan
University of Queensland†

June 21, 2012

**Abstract**

A Bayesian decision maker does not know which of several parameters is true. In each period she chooses an action 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. For the central economic models (systems of markets, noncooperative games) the relevance of equilibrium (market clearing, Nash equilibrium) is to a greater or lesser extent justified by a presumption

*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. We have benefitted from remarks by seminar participants at the University of Queensland and the 2012 Shanghai Microeconomics Workshop.

that we are now observing a steady state that emerged from a process of learning and adjustment. A natural and relevant 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, with positive probability, a rational decision maker's chosen actions become less and less informative, converging to an action that is completely uninformative, while the sequence of beliefs converges to a limiting belief with residual uncertainty about an unknown parameter. This parameter is chosen by nature before the first period. The decision maker begins each period with a belief about it, chooses an action, and observes an outcome. There is a known relationship between the parameter, the action, and the distribution of outcomes, and the next period's belief is given by Bayesian updating. In each period the decision maker receives a reward that can depend on the action and outcome or, more generally, on the action and the current belief, and her objective is to maximize the expectation of the discounted sum of rewards.

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 function of the current belief, and this policy dictates that the uninformative action $a^*$ is chosen at a critical belief $\omega^*$. Since the policy is continuous, the amount of experimentation is small when the prior belief is near $\omega^*$, 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. Thus there is a probabilistic ratchet effect drawing the sequence of beliefs toward $\omega^*$; more formally, the stochastic process given by the logarithm of the distance from the belief in period $t$ to $\omega^*$ is a supermartingale near $\omega^*$.

The criterion for a positive probability of convergence to $\omega^*$ depends on the first derivative of the policy function at $\omega^*$. We give conditions under which both the optimal policy and its first derivative varies continuously as the discount factor is varied near 0. Consequently policies allowing such asymptotically incomplete learning can be optimal even when the decision maker is not myopic and values experimentation.

Some of the oldest literature 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. Bergemann and Välimäki (2008) provide a recent survey and summary.

Multiagent environments with observational learning are studied by Banerjee (1992)

and Bikhchandani et al. (1992), and numerous subsequent papers. In such models failure to learn, and other inefficiencies, result from the fact that agents are not compensated for the external benefits of experimentation. Strulovici (2010) studies such externalities in an environment in which voter learns from experience about the benefits of various policies that are chosen in repeated elections. Multiagent learning models have been applied to political economy (e.g., Piketty (1995)) regulation and taxation (e.g., Laslier et al. (2003) and Berentsen et al. (2008)), pricing in industrial organization (e.g., Bergemann and Välimäki (1996), Harrison et al. (2011)), product design (e.g., Callander (2011)), behavioral economics (e.g., Ali (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 a single decision maker 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 the 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 also easily happen that the myopically optimal policy does not allow the sequence of beliefs to 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. We will see that the framework presented here has the results of the earlier paper as special cases.

The remainder has the following organization. Section 2 presents the general model of learning as a 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 provides conditions on the policy function under which there is a positive probability that the sequence of beliefs converges to a belief at which the action prescribed by the policy is uninformative.

Section 3 provides a series of computational results that are related to whether the condition identified in our first main result holds. These have the consequence that examples of this phenomenon exist for any number of possible values of the unknown parameter. A related issue is an analysis of the extent to which adopting a policy of less aggressive experimentation increases the likelihood of incomplete learning.

Section 4 proves the first main result by developing a version of the law of large numbers that formalizes the intuition described above concerning the logarithm of the distance from the current belief to $\omega^*$. 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 criterion for a positive probability of convergence given by the first result requires that the critical belief (the one for which the uninformative action is myopically optimal) is mapped to the uninformative action by the policy, and that the derivative of the policy function at that point satisfies 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. Our second main result provides 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.

An interesting and perhaps somewhat novel aspect of the argument is the use made

of the fact that certain operators are Lipschitz functions, where either the domain or the range (or both) is endowed with the metric that is usually used to induce the standard $C^r$ topology on the space of $C^r$ real valued functions on the space. Blume et al. (1982) study a somewhat different stochastic dynamic programming model, achieving $C^r$ value functions and policy functions by means of the implicit function theorem for Banach spaces; here we use the Lipschitz condition to achieve a favorable modulus of contraction, so our methods are somewhat different, but clearly closely related.

The third main result combines the first two, straightforwardly, to deliver the key conceptual conclusion: even an agent with a positive discount factor can have a positive probability of begin drawn into a Bayesian learning trap. It is described informally at the end of Section 2, then stated and proved in Section 6. Among its hypotheses is the requirement that the space of beliefs and the space of actions have the same dimension. (This allows the theory of the topological degree to be used to show that a policy function that is near the myopically optimal policy also maps some belief to the uninformative action.) Section 7 discusses ways in which this assumption might be relaxed, and other possible directions of generalization and extension, thereby concluding the paper.

# 2 The 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 $\delta_s$ 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 $\sigma$-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 measures with full support, i.e., those $\sigma \in \Delta(S)$ such that $\sigma(s) > 0$ for all $s$.

Let $\Theta$ be a finite set of possible values of a *parameter* $\tilde{\theta}$ that the decision maker learns 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 a stochastic process.

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 the probability $q_\theta(y|a)$ that $y$ is observed when $\theta$ is the parameter and $a$ is

chosen. We always assume that for each $\theta$ 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 $t$. When the prior belief is $\omega$ and action $a$ is chosen, there is a distribution $B(\omega, a) \in \Delta(\Omega)$ of the posterior belief given by

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

A general property of Bayesian updating is that the expectation of the posterior is the prior: $\mathbf{E}(\beta(\omega, a, y)|\omega, a) = \omega$. 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 $\tilde{\omega}_t$.

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

$$\pi : \Omega \to A.$$

That is, for all $t$, it is almost surely the case that $\tilde{a}_t = \pi(\tilde{\omega}_t)$. In this case $B(\tilde{\omega}_t, \pi(\tilde{\omega}_t))$ 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 always assume $\pi$ is continuous, and for the most part this function will also be $C^1$.

We say that *learning is asymptotically incomplete* if $\{\tilde{\omega}_t\}$ does not converge to a point in $\{\, \delta_\theta : \theta \in \Theta \,\}$. 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 $\omega^*$ at which $\tilde{\theta}$ is not known with certainty. We will always assume that $\omega^* \in \Delta^\circ(\Theta)$; Section 7 has some speculative remarks concerning the possibility that the support of $\omega^*$ is neither a singleton nor all of $\Theta$.

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 $\omega$ and $y$, so that $B(\omega, a^*) = \delta_\omega$. If $\pi(\omega^*)$ is uninformative and $\tilde{\omega}_t = \omega^*$, then it is almost surely the case that $\tilde{\omega}_s = \omega^*$ for all $s \geq t$. In most settings there is no reason to expect that $\tilde{\omega}_0 = \omega^*$, and examples in which there is a positive probability of a sequence of events leading to $\tilde{\omega}_t = \omega^*$ for some $t$ are similarly quite 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. Since $\pi$ is continuous, the function $\omega \mapsto B(\omega, \pi(\omega))$ is continuous, so there cannot be a positive probability of convergence to an $\omega^* \in \Delta^\circ(\Theta)$ unless $a^* = \pi(\omega^*)$ is uninformative, as we assume henceforth.

There are now two main questions:

1) Under what conditions on $\pi$ 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 $\delta$ is positive and sufficiently small?

The case of $\delta$ close to one is also interesting, but easily answered. For $\delta$ sufficiently close to one it cannot be optimal to choose $a^*$ because if it was optimal to do so in a single period, it would be optimal to do so in every subsequent period, but the value in future periods of the information gleaned from a highly informative action exceeds the loses in the current period. (A related result is Theorem 4.3 of Aghion et al. (1991), which asserts that as $\delta \to 1$, the discounted per period value of the problem approaches the discounted per period value of a decision maker who sees $\theta$.) Since the graph of the best response correspondence is upper semicontinuous, it follows that for $\delta$ close to one there is a neighborhood of $a^*$ that does not contain an optimal action for any belief.

**Remark:** Consider the possibility that, as in McLennan (1984), $\Omega$ 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 $\omega^*$ be a belief such that $\pi(\omega^*) = a^*$. Suppose that $\tilde{\omega}_0 = \omega_0$ almost surely, where $\omega_0$ is between $\delta_{\theta_1}$ and $\omega^*$. It can easily happen that for all $t$, $\tilde{\omega}_t$ is almost surely between $\delta_{\theta_1}$ and $\omega^*$ because the action prescribed by $\pi$ 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 $\delta_{\theta_1}$ or $\omega^*$, 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 $\pi_\delta$ 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 $\pi_0$ has all the features described above. McLennan (1984) showed that it can happen that for all $\delta$ in some interval $[0, \overline{\delta})$, $\pi_\delta$ also has all of these features.

The mechanism leading to asymptotic incomplete learning presented here is more general and robust. It is not limited to the case of two possible parameters, nor does it depend on there being an impenetrable barrier to learning 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.

Of course $\tilde{\omega}_t \to \omega^*$ if and only if $\ln \|\tilde{\omega}_t - \omega^*\| \to -\infty$. (Here $\|\cdot\|$ may be *any* norm on $\mathbb{R}^\Theta$.) Our 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 $\omega^*$. When $\Omega$ is one dimensional and the process $\{\tilde{\omega}_t\}$ is confined one of the two intervals determined by $\omega^*$, 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

$$H_0 = \{\, \tau \in \mathbb{R}^\Theta : \sum_\theta \tau_\theta = 0 \,\} \quad \text{and} \quad H_1 = \{\, \tau \in \mathbb{R}^\Theta : \sum_\theta \tau_\theta = 1 \,\}.$$

That is, $H_1$ is the hyperplane in $\mathbb{R}^\Theta$ that contains $\Omega$, and $H_0$ is the parallel hyperplane through the origin. Let

$$S = \{\, \sigma \in H_0 : \|\sigma\| = 1 \,\}$$

be the unit sphere in $H_0$. Any $\omega \in H_1$ has a representation of the form $\omega = \omega^* + r\sigma$ where $\sigma \in S$ and $r \geq 0$, and this representation is unique if $\omega \neq \omega^*$.

Let $\kappa : S \times [0, \infty) \times Y \to \Omega$ be the function

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

From this point forward we assume that $\pi$ and the functions $q_\theta(y|\cdot)$ are $C^1$. Elementary calculus implies that $\frac{\partial \kappa}{\partial r}$ is a well defined continuous function from $S \times [0, \infty) \times Y$ to $H_0$. We define $\nu : S \times [0, \infty) \times Y \to H_1$ by setting

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

**Lemma 1.** *$\nu$ is continuous.*

*Proof.* Clearly $\nu$ is continuous at every $(\sigma, r)$ with $r > 0$, and the restriction of $\nu$ to $S \times \{0\}$ is also continuous. If $\{(\sigma_n, r_n)\}$ is a sequence of points with $r_n > 0$ 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 $\nu(\sigma_n, r_n, y) = \frac{\partial \kappa}{\partial r}(\sigma_n, r'_n, y) - \sigma$, so the continuity of $\frac{\partial \kappa}{\partial r}$ implies that $\nu(\sigma_n, t_n, y) \to \nu(\sigma, 0, y)$. □

We define a function $B_\pi : S \times [0,\infty) \to \Delta(H_1)$ by letting $B_\pi(\sigma, r)$ be the element of $\Delta(H_1)$ that assigns probability

$$\sum_\theta (\omega^*_\theta + r\sigma_\theta) \sum_{y\,:\,\omega^*+\sigma+\nu(\sigma,r,y)\in E} q_\theta(y|\pi(\omega^* + r\sigma))$$

to each event $E \subset \Delta(H_1)$. Clearly $B_\pi$ is continuous. Let $q(y|a^*)$ be the common value of $q_\theta(y|a^*)$, which is the probability (for any $\theta$) of observing $y$ when $a^*$ is chosen. Roughly speaking, $B_\pi(\sigma, 0)$ assigns probability $q(y|a^*)$ to each $\omega^* + \sigma + \nu(\sigma, 0, y)$.

Our first main result is as follows. The proof is given at the end of Section 4, after the supporting statistical result has been developed.

**Theorem 1.** *If, for all $\sigma \in S$,*

$$\int_{H_1} \ln \|\omega - \omega^*\| \, dB_\pi(\sigma, 0) = \sum_{y\in Y} q(y|a^*) \ln \|\sigma + \nu(\sigma, 0, y)\| < 0, \qquad (*)$$

*then for any sufficiently small neighborhood $U$ of $\omega^*$, if the probability that $\tilde\omega_0 \in U$ is positive, then there is a positive probability that $\tilde\omega_t \to \omega^*$.*

The expectation of $\ln \|\sigma + \nu(\sigma, 0, y)\|$ will be negative, due to the concavity of the logarithm function, if $\nu(\sigma, 0, y)$ is predominantly parallel to $\sigma$, as is necessarily the case when $\Omega$ is 1-dimensional, and its norm is not too large. On the other hand the expectation will be positive if $\nu(\sigma, 0, y)$ is predominantly orthogonal to $\sigma$. Thus there arises the question of whether $(*)$ can hold for all $\sigma$, which we address in the next section.

A natural intuition is that the stochastic process on $S$ given by $B_\pi(\cdot, 0)$ will typically be ergodic, which is to say that it has a unique invariant distribution $\mu$, and, regardless of the initial state, the long run empirical distribution will converge to it. One might then expect that the sequence of beliefs will be drawn to $\omega^*$ or repelled from it according to whether

$$\int_S \Big( \int_{H_1} \ln \|\omega - \omega^*\| \, dB_\pi(\sigma, 0) \Big) \, d\mu < 0.$$

Unfortunately, since the support of $B_\pi(\sigma, 0)$ is always finite, it is hard to state checkable conditions on $B_\pi(\cdot, 0)$ that guarantee ergodicity. Among other things, it is hard to know whether there is a finite $S' \subset S$ such that for all $\sigma \in S'$, the support of $B_\pi(\sigma, 0)$ is contained in $S'$. In addition, even if $B_\pi(\cdot, 0)$ induces an ergodic process, the process $\tilde\sigma_t = (\tilde\omega_t - \omega^*)/\|\tilde\omega_t - \omega^*\|$ may fail to behave in a similar manner when $\tilde r_t = \|\tilde\omega_t - \omega^*\|$ is closed to zero because of such "hidden invariant sets" or other anamolies.

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 6, has the following intuition: under natural and easily verified conditions, the optimal policy $\pi_\delta$, and its derivative, will vary continuous as $\delta$ varies in a neighborhood of zero. Provided that the dimension of $\Omega$ is the same as the dimension of $A$, it follows from the theory of the topological degree that for small $\delta > 0$, some point near $\omega^*$ is mapped to $a^*$. Since choosing $a^*$ minimizes learning, it cannot be optimal to do so for positive $\delta$ unless it is also myopically optimal, so we have $\pi_\delta(\omega^*) = a^*$ for small positive $\delta$. Since the derivative of $\pi$ varies continuously with $\delta$, if $(*)$ holds for all $\sigma \in S$ when $\pi = \pi_0$, then it also holds for all $\sigma$ when $\pi = \pi_\delta$ and $\delta > 0$ is sufficiently small.

# 3 Computation and Examples

From a computational point of view, the expectation of the logarithm of the norm of a random vector is cumbersome. In this section we consider how the right hand side of $(*)$ varies as we decrease the intensity of experimentation. This is interesting in itself, and in addition it allows us to analyze the right hand side of $(*)$ in terms of its first and second derivative with respect to the intensity of experimentation. One consequence will be a wealth of examples in which $(*)$ is satisfied for all $\sigma$.

We retain the framework of the last section: $\pi : \Omega \to A$ is a $C^1$ policy, $\omega^*$ is a point in the interior of $\Omega$, and $\pi(\omega^*) = a^*$ is uninformative. For $\sigma \in S$ and $y \in Y$ define vectors $\rho(\sigma, 0, y) \in \mathbb{R}^\Theta$ by setting

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

and

**Lemma 2.** *For $\sigma \in S$ and $y \in Y$ we have*

$$\nu_\theta(\sigma, 0, y) = \rho_\theta(\sigma, 0, y) - \omega^*_\theta \sum_{\theta' \in \Theta} \rho_{\theta'}(\sigma, 0, y).$$

*Proof.* Set

$$\gamma_\theta(\sigma, r, y) = (\omega^*_\theta + r\sigma_\theta) q_\theta(y|\pi(\omega^* + r\sigma)).$$

Then

$$\beta_\theta(\omega^* + r\sigma, \pi(\omega^* + r\sigma), y) = \frac{\gamma_\theta(\sigma, r, y)}{\sum_{\theta'} \gamma_{\theta'}(\sigma, r, y)}.$$

Noting that $\gamma_\theta(\sigma, 0, y) = \omega^*_\theta q(y|a^*)$ and $\sum_{\theta'} \gamma_{\theta'}(\sigma, 0, y) = \sum_{\theta'} \omega^*_{\theta'} q(y|a^*) = q(y|a^*)$, we use elementary calculus to compute that

$$\frac{\partial \kappa_\theta}{\partial r}(\sigma, 0, y) = \frac{\frac{\partial \gamma_\theta}{\partial r}(\sigma, 0, y) q(y|a^*) - \gamma_\theta(\sigma, 0, y) \sum_{\theta'} \frac{\partial \gamma_{\theta'}}{\partial r}(\sigma, 0, y)}{q(y|a^*)^2}$$

$$= \frac{1}{q(y|a^*)} \big( \frac{\partial \gamma_\theta}{\partial r}(\sigma, 0, y) - \omega^*_\theta \sum_{\theta'} \frac{\partial \gamma_{\theta'}}{\partial r}(\sigma, 0, y) \big).$$

In turn we have

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

Using the fact that $\sum_{\theta'} \sigma_{\theta'} = 0$, we find that

$$\sigma_\theta + \nu_\theta(\sigma, 0, y) = \frac{\partial \kappa_\theta}{\partial r}(\sigma, 0, y) = \sigma_\theta + \rho_\theta(\sigma, 0, y) - \omega^*_\theta \sum_{\theta' \in \Theta} \rho_{\theta'}(\sigma, 0, y),$$

as desired. □

We now analyze the extent to which increasing the amount of experimentation affects the probability of falling into the learning trap. We now parameterize the key variables from Section 2 by a number $\alpha \geq 0$, setting

$$\pi^\alpha : \omega \to a^* + \alpha(\pi(\omega) - a^*).$$

Evidently $\alpha$ measures the aggressiveness of experimentation of $\pi^\alpha$ near $\omega^*$. Let

$$\rho^\alpha_\theta(\sigma, 0, y) = \frac{1}{q(y|a^*)} \cdot \frac{\partial q_\theta(y|\pi^\alpha(\omega^* + r\sigma))}{\partial r}\Big|_{r=0}$$

and

$$\nu^\alpha_\theta(\sigma, 0, y) = \rho^\alpha_\theta(\sigma, 0, y) - \omega^*_\theta \sum_{\theta' \in \Theta} \rho^\alpha_{\theta'}(\sigma, 0, y).$$

The chain rule gives $\rho^\alpha(\sigma, 0, y) = \alpha\rho(\sigma, 0, y)$, so $\nu^\alpha(\sigma, 0, y) = \alpha\nu(\sigma, 0, y)$. Therefore

$$\int_{H_1} \ln \|\omega - \omega^*\| \, dB_{\pi^\alpha}(\sigma, 0) = \sum_{y \in Y} q(y|a^*) \ln \|\sigma + \alpha\nu(\sigma, 0, y)\|.$$

As we mentioned earlier, the results so far pertain to any norm $\|\cdot\|$ on $\mathbb{R}^\Theta$, but applications of the following result require that we assume that $\|\cdot\|$ is derived from an inner product $\langle \cdot, \cdot \rangle$. The proof is a matter of elementary calculus, and is omitted.

**Lemma 3.** *For vectors $v$ and $w$ in an inner product space with $v \neq 0$, if $g(s) = \ln \|v + sw\|$, then*

$$g'(0) = \frac{\langle v, w\rangle}{\|v\|^2} \quad \textit{and} \quad g''(0) = \frac{\|v\|^2\|w\|^2 - 2\langle v, w\rangle^2}{\|v\|^4}.$$

We now have

$$\frac{d\big(\int_{H_1} \ln\|\omega - \omega^*\|\, dB_{\pi^\alpha}(\sigma, 0)\big)}{d\alpha}\Big|_{\alpha=0} = \sum_y q(y|a^*)\frac{\langle \sigma, \nu(\sigma, 0, y)\rangle}{\|\sigma\|^2} = 0$$

and

$$\frac{d^2\big(\int_{H_1} \ln\|\omega - \omega^*\|\, dB_{\pi^\alpha}(\sigma, 0)\big)}{d\alpha^2}\Big|_{\alpha=0} = \sum_y q(y|a^*)\frac{\|\sigma\|^2\|\nu(\sigma, 0, y)\|^2 - 2\langle \sigma, \nu(\sigma, 0, y)\rangle^2}{\|\sigma\|^4}. \tag{$*$}$$

This finding suggests that when there is already a small amount of experimentation as one moves away from $\omega^*$, the effect of further reducing experimentation, by replacing $\pi$ with $\pi^\alpha$ where $\alpha < 1$, is primarily to slow the process down, with little consequence for the probability of eventual convergence to $\omega^*$. 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 $\alpha^2$; 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 $\Omega$ and $A$ are both 1-dimensional, the right hand side of $(*)$ reduces to $-\sum_y q(y|a^*)\|\nu(\sigma, 0, y)\|^2$ because $\nu(\sigma, 0, y)$ is necessarily a scalar multiple of $\sigma$, so in this case there is necessarily a positive probability of convergence to $\omega^*$ when $\alpha$ is sufficiently small.

We now look at the linear case:

$$q_\theta(y|\pi(\omega^* + \tau)) = q(y|a^*) + \langle \tau, c_{\theta,y}\rangle,$$

Here the vector $c_{\theta,y}$ is a proxy for the derivative of $q_\theta(y|\pi(\cdot))$, so we can imagine that this formula holds, in the first order approximate sense, only for $\tau$ near zero.. Since we only care about the inner product of these vectors with elements of $H_0$, they may be taken to be in $H_0$, and are otherwise arbitrary except for the requirement that the total probability assigned to the various $y$ is always one, which means that for each $\theta$ we have $\sum_y c_{\theta,y} = 0$. We compute that

$$\rho_\theta(\sigma, 0, y) = \frac{1}{q(y|a^*)}\langle \sigma, c_{\theta,y}\rangle$$

and

$$\nu_\theta(\sigma, 0, y) = \frac{1}{q(y|a^*)}\big(\langle \sigma, c_{\theta,y}\rangle - \omega^*_\theta \sum_{\theta'} \langle \sigma, c_{\theta',y}\rangle\big) = \frac{1}{q(y|a^*)}\langle \sigma, d_{\theta,y}\rangle$$

where $d_{\theta,y} = c_{\theta,y} - \omega^*_\theta \sum_{\theta'} c_{\theta',y}$. Evidently $\sum_y d_{\theta,y} = 0$, so replacing the $c_{\theta,y}$ with the $d_{\theta,y}$ gives another linear system that results in the same vectors $\nu(\sigma, 0, y)$. One can easily compute that $\sum_\theta d_{\theta,y} = 0$. The upshot of this is that in studying the possibilities for the $\nu(\sigma, 0, y)$ there is no loss of generality in assuming that for each $y$, $\sum_\theta c_{\theta,y} = 0$. In this way we arrive at the formula

$$\nu_\theta(\sigma, 0, y) = \frac{1}{q(y|a^*)}\langle \sigma, c_{\theta,y}\rangle.$$

For each $y$ let $\ell_y : H_0 \to H_0$ be the linear function with component functions

$$\ell_{\theta,y}(\sigma) = \langle \sigma, c_{\theta,y}\rangle.$$

(Since $\sum_\theta c_{\theta,y} = 0$, the image of $\ell_y$ is contained in $H_0$.) The condition $\sum_y c_{\theta,y} = 0$ implies that $\sum_y \ell_y = 0$. Aside from this constraint, these functions are arbitrary. We can now compute that

$$\|\nu(\sigma, 0, y)\|^2 = \frac{1}{q(y|a^*)^2}\|\ell_y(\sigma)\|^2,$$

$$\langle \sigma, \nu(\sigma, 0, y)\rangle = \frac{1}{q(y|a^*)}\langle \ell_y(\sigma), \sigma\rangle,$$

and

$$\frac{d^2\big(\int_{H_1} \ln\|\omega - \omega^*\|\, dB_{\pi^\alpha}(\sigma, 0)\big)}{d\alpha^2}\Big|_{\alpha=0} = \sum_y \frac{1}{q(y|a^*)} \frac{\|\sigma\|^2\|\ell_y(\sigma)\|^2 - 2\langle \sigma, \ell_y(\sigma)\rangle^2}{\|\sigma\|^4}.$$

In order to show that Bayesian learning traps can, in fact, occur, it suffices to provide a system of linear functions $\{\ell_y : H_0 \to H_0\}_{y\in Y}$ with $\sum_y \ell_y = 0$ such that the right hand side is negative for all nonzero $\sigma \in H_0$. But this is easy: for example, one can take $Y = \{y_1, y_2\}$, $\ell_{y_1}(\sigma) = \sigma$, and $\ell_{y_2}(\sigma) = -\sigma$.

# 4 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 $\Omega$. We study a stationary Markov process $\tilde\omega_0, \tilde\omega_1, \tilde\omega_2, \ldots$ in $\Omega$ 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

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

Let $\ell : \Omega \to \mathbb{R}$ be a measurable function. We study conditions on $P$ and $\ell$ 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 4.** *Let $\tilde{x}$ be a random variable with cumulative distribution function $\Phi$. 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*

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

*for 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 $\Phi_k$, then $\mathbf{E}(\tilde{x}_k) < 0$.*

*(b) for each $\omega \in \Omega$ such that $\ell(\omega) < 0$ there is some $k$ such that for all $x \in \mathbb{R}$,*

$$P\big(\{\omega' \in \Omega : \ell(\omega') \geq \ell(\omega) + x\} | \omega\big) \leq 1 - \Phi_k(x).$$

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

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

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

*Proof.* In view of Lemma 4, 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 $y \le 0$. For each $T = 0, 1, 2, \ldots$ let $p_T : \Omega \to [0, 1]$ be the function

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

It suffices to show that for a given $\omega_0$ such that $\ell(\omega_0) < 0$ we have $p_T(\omega_0) \le Ce^{\beta\ell(\omega_0)}$ 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 $P(\{\omega' \in \Omega : \ell(\omega') \ge \ell(\omega) + x\}|\omega) \le 1 - \Phi_k(x)$ for all $x \in \mathbb{R}$. Then

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

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

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

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

*Proof of Theorem 1.* For each $\sigma \in S$, if $\omega$ is distributed according to $B_\pi(\sigma, 0)$, then the expectation of $\ln\|\omega - \omega^*\|$ is negative. Since $B_\pi(\sigma, 0)$ 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 $(\sigma', r')$ in some neighborhood of $(\sigma, 0)$ in $S \times [0, \infty)$, this distribution also first order stochastically dominates the distribution of $\ln\|\omega - \omega^*\|$ when $\omega$ is distributed according to $B_\pi(\sigma', r')$. 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. □

# 5 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 straightforward extenions of definitions that are standard (cf. Hirsch (1976)) for manifolds with boundary.

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 $\{\varphi_g : U_g \to \mathbb{R}^m\}_{g \in G}$ where $\{U_g\}$ is an open cover of $M$, each $\varphi_g$ 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 $\{\varphi_g : U_g \to \mathbb{R}^m\}_{g \in G}$ 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 $\mathbb{R}^m_{\geq} = [0, \infty)^m$ 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 $\varphi_g$ 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 $C^\infty$ atlas for $\Omega$ 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

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

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.

Suppose that $\{\varphi_g : U_g \to \mathbb{R}^m_{\geq}\}_{g \in G}$ is a $C^r$ atlas for $M$ and $\{K_g\}$ is a locally finite collection of compact differentiation domains whose interiors cover $M$, with $K_g \subset U_g$ for

all $g$. For $f, f' \in C^s(M)$ let

$$d_s^g(f,f') = \max \left| \frac{\partial^t (f \circ \varphi_g^{-1})}{\partial x_{i_1} \cdots \partial x_{i_t}}(\varphi_g(p)) - \frac{\partial^t (f' \circ \varphi_g^{-1})}{\partial x_{i_1} \cdots \partial x_{i_t}}(\varphi_g(p)) \right|$$

where the maximum is over all $p \in K_g$, $t = 0, \ldots, s$, and $i_1, \ldots, i_t = 1, \ldots, m$. It is easy to verify that $d_s^g$ is a metric. The *strong $C^s$ topology* on $C^s(M)$ is the topology that has a basis of open sets of the form $\{ f' : d_s^g(f,f') < \varepsilon_g \}$ where $f \in C^s(M)$ and $\{\varepsilon_g\}_{g \in G}$ is a system of positive numbers. For any differentiation domain $D \subset M$ the strong $C^s$ topology on $C^s(D)$ is the quotient topology induced by the map $f \mapsto f|_D$.

If $D$ is compact, it is covered by the interiors of some finite collection $K_{g_1}, \ldots, K_{g_H}$. In this case the topology of $C^s(D)$ is induced by the metric

$$d_s^D(f,f') = \max_{h=1,\ldots,H} d_s^{g_h}(f,f').$$

Note that $d_s^D$ is the metric derived from the norm

$$\|f\| = \max_{h,t,i_1,\ldots,i_t,p} \left| \frac{\partial^t (f \circ \varphi_{g_h}^{-1})}{\partial x_{i_1} \cdots \partial x_{i_t}}(\varphi_{g_h}(p)) \right|.$$

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.

An unusual and interesting aspect of our methods is that they depend not just on the topology induced by $d_s^D$, but on the metric itself. Recall that for general metric spaces $(X,d)$ and $(Y,e)$ a function $f : X \to Y$ is *Lipschitz* if there is a $\Lambda > 0$ such that $e(f(x'), f(x'')) \leq \Lambda d(x', x'')$ for all $x', x'' \in X$. It is *Lipschitz at $x \in X$* if there is a neighborhood $U$ of $x$ such that $f|_U$ is Lipschitz, and it is *locally Lipschitz* if it is Lipschitz at each point in $X$. Of course the metric $d_s^D$ depends on the atlas and the sets $K_{g_1}, \ldots, K_{g_H}$, so in order to have a meaningful notion of what it means for a function to or from $C^s(D)$ to be locally Lipschitz, the definition 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 $d_s^D$ also denote the metric given by

$$d_s^D(f,f') = d_s^D(f_1,f_1') + \cdots + d_s^D(f_n,f_n').$$

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 $C^0(D, U)$.

# 6 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 $\delta$ varies in a neighborhood of $\delta = 0$.

Initially we let the space of *states* $\Omega$ and the space of *actions* $A$ be metric spaces, with $\Omega$ compact. Let $u : \Omega \times A \to \mathbb{R}$ and $P : \Omega \times A \to \Delta(\Omega)$ be continuous *reward* and *transition* functions. 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 $P(\tilde{\omega}_t, \tilde{a}_t)$. A *policy* for this problem is a sequence $\pi_0, \pi_1, \ldots$ of measurable functions $\pi_t$ taking histories $(\tilde{\omega}_0, \tilde{a}_0), \ldots, (\tilde{\omega}_{t-1}, \tilde{a}_{t-1}), \tilde{\omega}_t$ to action choices in period $t$. A policy and an initial state $\tilde{\omega}_0$ induce a probability measure on infinite histories $(\tilde{\omega}_0, \tilde{a}_0), (\tilde{\omega}_1, \tilde{a}_1), \ldots$.

Our methods require that we restrict attention to a compact $D \subset \Omega \times A$. For each $\omega \in \Omega$ let $D(\omega)$ be the set of $a$ such that $(\omega, a) \in D$, which we assume is nonempty. We say that $D$ *shields the optimizers of* $u$ if there is an $\varepsilon > 0$ such that

$$u(\omega, a) < \max_{a' \in A} u(\omega, a') - \varepsilon$$

for all $(\omega, a) \in (\Omega \times A) \setminus D$. In this circumstance the value function is bounded above, and the agent can always prevent the reward in period $t$ from falling below a lower bound by choosing an action in $D(\tilde{\omega}_t)$. Therefore for each initial state $\omega_0$ the supremum, over all policies, of the expectation of the sum of discounted rewards when $\tilde{\omega}_0 = \omega_0$ is a well defined finite number. This is the *value* of $\omega_0$, and the *value function* for the problem is the function $V_\delta : \Omega \to \mathbb{R}$ taking each state in $\Omega$ to its value.

We apply the standard methodology of dynamic programming, by analyzing the value of the problem as a function of $\omega_0$. There is an operator $J : C(D) \to C(\Omega)$ that is defined by setting

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

There is a second operator

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

For $\delta \in \mathbb{R}$ let $L_\delta : C(\Omega) \to C(\Omega)$ be the operator

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

**Lemma 5.** *If $D$ shields the optimizers of $u$, then there is a $\overline{\delta} > 0$ such that for all $\delta \in [0, \overline{\delta})$, $V_\delta$ is the unique fixed point of $L_\delta$. If $A$ is compact, then one may take $\overline{\delta} = 1$.*

*Proof.* It is easy to show that for $\delta$ sufficiently small actions outside of $D(\omega)$ cannot be optimal when the state is $\omega$. That is, the given problem has the same value as the problem in which the action at $\omega$ is required to be in $D(\omega)$. For that problem, and when $A$ is compact, the result is standard; Maitra (1968) is a suitable reference. □

From now on we assume that $\Omega$ is a compact $m$-dimensional $C^2$ manifold with corners, and that $A$ is a $C^2$ manifold without boundary. Then $A$ is 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. We say that $u : \Omega \times A \to \mathbb{R}$ satisfies the *standard conditions* if it is $C^2$, with Lipschitz second order derivatives, and, for each $\omega \in \Omega$, there is a unique maximizer $\pi_u(\omega)$ of $u(\omega, \cdot)$ at which the second order necessary conditions for maximization hold strictly.

The function $P$ is said to be *second order smoothing near* $V \in C^2(\Omega)$ if there a neighborhood $Z \subset C^2(\Omega)$ of $V$ such that $K(Z) \subset C^2(\Omega \times A)$ and for any compact differentiation domain $D$ the function $V' \mapsto K(V')|_D$ from $W$ to $C^2(D)$ is Lipschitz. We do not provide methods for checking that this condition holds; in the problem of interest to us this is easy, and this will be the case for many other applications.

**Theorem 2.** *If $u$ satisfies the standard conditions, $D \subset \Omega \times A$ is a compact differentiation domain that shields the optimizers of $u$, and $P$ is second order smoothing near $V_0$, then there is a $\overline{\delta} > 0$ such that for all $\delta \in (-\overline{\delta}, \overline{\delta})$ the optimal policy $\pi_\delta : \Omega \to A$ is $C^1$ and the value function $V_\delta$ is $C^2$. In addition, the maps $\delta \mapsto \pi_\delta$ and $\delta \mapsto V_\delta$ from $(-\overline{\delta}, \overline{\delta})$ to $C^1(D, A)$ and $C^2(\Omega)$ are Lipschitz.*

In preparation for the proof we study how $J(u)$ and the optimal policy vary when we perturb $u$ in $C^2(D)$.

**Proposition 2.** *If $u$ satisfies the standard conditions and $D$ shields the optimizers of $u$, then there is a neighborhood $W \subset C^2(D)$ of $u$ such that:*

*(a) For each $u' \in W$ and each $\omega$ there is a unique maximizer $\pi_{u'}(\omega)$ 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 $u' \in W$, $\pi_{u'} : \Omega \to A$ is $C^1$.*

*(c) The operator $u' \mapsto \pi_{u'}$ 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 $\pi_u(\omega)$, 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 $u' \mapsto \pi_{u'}$ 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 $\Omega = [0, 1]$ and $A = (0, 1)$; it will be clear that all the steps in the argument generalize in a straightforward manner. Fully differentiating the equation

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

and rearranging gives

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

We can now decompose the difference between $\frac{d\pi_u}{d\omega}$ and $\frac{d\pi_{u'}}{d\omega}$ into two parts: i) the consequence of replacing $\pi_u(\omega)$ with $\pi_{u'}(\omega)$ 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 $d_0^{(\Omega,A)}(\pi_{u'}, \pi_u)$ 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

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

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

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 $\frac{d\pi_{u'}}{d\omega}$ 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 $\pi_u$ with $\pi_{u'}$ and the effects of replacing $u$ with $u'$ establishes (e). □

*Proof of Theorem 2.* After replacing $W$ from Proposition 2 with a smaller neighborhood of $u$, we may assume that $J(W) \subset Z$. Since $d_2^\Omega$ and $d_2^D$ respect multiplication by scalars, there is a $\overline{\delta} > 0$ and a constant $\lambda \in (0,1)$ such that for all $\delta \in [0, \overline{\delta}]$, $L_\delta$ is a contraction with modulus of contraction at most $\lambda$. Since $C^2(D)$ is a Banach space, and consequently complete, in this circumstance $L_\delta$ must have a fixed point $V_\delta$. Let $\pi_\delta$ be the optimal policy function for the discounted problem with discount factor $\delta$. A standard argument based on iteratively applying the operator $L_\delta$ shows that the map $\delta \mapsto V_\delta$ is Lipschitz with Lipschitz constant $1/(1-\lambda)$. In view of the last result, it follows that the map $\delta \to \pi_\delta$ is also Lipschitz. □

We can now state and prove the third main result.

**Theorem 3.** *In addition to the assumptions of Theorem 1, suppose that $u : \Omega \times A \to \mathbb{R}$ satisfies the standard conditions, $D \subset \Omega \times A$ is a compact differentiation domain that shields the optimizers of $u$, $\pi = \pi_0$ is the myopically optimal policy (that is, $u(\pi_0(\omega)) = \max_{a \in A} u(\omega, a)$) and the functions $q_\theta(y|\cdot)$ are $C^2$. Finally, assume that the dimension of $\Omega$ is the same as the dimension of $A$, and that the derivative of $\pi_0$ at $\omega^*$ is nonsingular. Then there is a $\overline{\delta} > 0$ such that for all $\delta \in (-\overline{\delta}, \overline{\delta})$, if $\pi_\delta$ is the optimal policy for the problem with discount factor $\delta$, $\{\tilde{\omega}_t\}$ is a stationary Markov process in which $B(\tilde{\omega}_t, \pi_\delta(\tilde{\omega}_t))$ is the distribution of $\tilde{\omega}_{t+1}$ conditional on $\tilde{\omega}_t$, and there is a positive probability that $\tilde{\omega}_0$ is in a sufficiently small neighborhood of $\omega^*$, then there is a positive probability that $\tilde{\omega}_t \to \omega^*$.*

*Proof.* Suppose that the conclusion of Theorem 2 holds: in some interval $[0, \overline{\delta}]$ the function $\delta \to \pi_\delta$ is continuous when the range has the $C^1$ topology. By the theory of the topological degree, continuity relative to the $C^0$ topology implies that for sufficiently small $\delta$ there is some point near $\omega^*$ that is mapped to $a^*$. Since $a^*$ minimizes learning, it cannot be optimal for small positive $\delta$ unless it is also myopically optimal, and since the derivative of $\pi_0$ at $\omega^*$ is nonsingular, it follows that we must have $\pi_\delta(\omega^*) = a^*$. It is easy to see that $B_{\pi_\delta}(\sigma, 0)$ is jointly continuous as a function of $\sigma$ and the derivative

of $\pi_\delta$ at $\Omega^*$, and the latter varies continuously with $\delta$, so for sufficiently small $\delta$ we have $\int_{H_1} \ln \|\omega - \omega^*\| \, dB_{\pi_\delta}(\sigma, 0) < 0$ for all $\sigma \in S$, after which the desired conclusion follows from Theorem 1.

The conclusion of Theorem 2 will follow once we verify its hypotheses, and for this it remains only to show that the relevant $P : \Omega \times A \to \Delta(\Omega)$ is second order smoothing near the value function of the myopic problem. Concretely we have

$$K(V)(\omega, a) = \sum_\theta \omega_\theta \sum_y q_\theta(y|a) V(\beta(\omega, y, a)).$$

Since the functions $q_\theta(y|\cdot)$ are $C^2$, so are the functions $\beta(\cdot, \cdot, y)$. The chain rule implies that $K(V)$ and its first and second derivatives depend in a continuous manner on $V$ and its first and second derivatives. □

# 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. We have given concrete methods to compute whether the relevant conditions hold, and shown that for any finite space of possible parameters, there exist examples in which these conditions actually hold.

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 $\Omega$, 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 $\Theta$. 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.

Ali, S. N. (2011). Learning self-control. *Quarterly Journal of Economics*, 126:857–893.

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.

Bergemann, D. and Välimäki, J. (1996). Learning and strategic pricing. *Econometrica*, 64:1125–1149.

Bergemann, D. and Välimäki, J. (2008). Bandit problems. In Durlauf, S. and Blume, L., editors, *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, New York.

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.

Blume, L., Easley, D., and O'Hara, M. (1982). Characterization of optimal plans for stochastic dynamic programs. *Journal of Economic Theory*, 28:221–234.

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.

Maitra, A. (1968). Discounted dynamic programming on compact metric spaces. *Sankhya:the Indian Journal of Statistics, Series A*, 30:211–216.

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. (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.

Strulovici, B. (2010). Learning while voting: Determinants of collective experimentation. *Econometrica*, 78:933–971.