Active Learning with Misspecified Beliefs[*]

Drew Fudenberg[†] Gleb Romanyuk[‡] Philipp Strack[§]

February 2, 2016

Abstract

We study learning and information acquisition by a Bayesian agent who is is misspecified in the sense that his prior belief assigns probability zero to the true state of the world. In our model, at each instant the agent takes an action and observes the corresponding payoff, which is the sum of the payoff generated by a fixed but unknown function and an additive error term. We provide a complete characterization of asymptotic actions and beliefs when the agent's subjective state space is a doubleton. A simple example with three actions shows that in a misspecified environment a myopic agent's beliefs converge while a sufficiently patient agent's beliefs do not. This shows that examples of myopic agents with non-converging beliefs in the prior literature require all actions to be informative if the agent is non-myopic, and illustrates a novel interaction between misspecification and the agent's subjective interest rate

1 Introduction

In many economic settings, agents are uncertain about the payoff consequences of their actions, and the action they choose influences both their current payoff and the information they receive. A fully myopic agent will ignore the value of future information, and so may repeatedly take actions that would not be optimal under full information. If the agent is not myopic, she may choose to experiment with actions that do not maximize their immediate expected return. The optimal "active learning" rule here for a Bayesian agent trades off the gains from experimentation, or "exploration", versus short run returns, or "exploitation". The details of these active learning rules have been extensively studied in the case where

[*] We are grateful for comments from Jerry Green, Bruno Strulovici, and Sergey Vorontsov.

[†] Harvard University.

[‡] Harvard University.

[§] UC Berkeley.

the agent is a Bayesian whose prior is rich enough to include the true state of the world. In many economic situations, however, it is plausible that the agent's prior is misspecified, in the sense that it assigns probability 0 to the true state of the world, because the the space of possible models is quite large (Diaconis and Freedman (1986)). We do not attempt to provide micro-foundations for why the agent possibly has misspecified models, but take the misspecification as given and characterize the resulting behavior.

This paper is the first study of active learning and information acquisition by a misspecified Bayesian. In our model, at each instant the agent takes an action and observes the corresponding payoff, which is the sum of the payoff generated by a fixed but unknown function and an additive error term whose distribution the agent knows corresponds to a Brownian motion. The agent thinks there are two possible payoff functions, yet the true payoff function is neither of these. We give a complete characterization of the limit behavior of beliefs and actions, and in particular we determine when beliefs converge to a steady state. The agent's beliefs do not converge if all steady states are "repelling" in the sense that an informative action played near the steady state generates the signals in favor of another steady state. Using this fact, we show that if in addition there is an uninformative action, and two informative ones, and the uninformative action is myopically optimal for intermediate beliefs, a myopic agent's beliefs will converge to a steady state, while a sufficiently patient agent's beliefs will oscillate indefinitely. Finally we characterize the long-run outcome in a one-armed bandit problem, and show how it depends on which of the subjectively possible payoff functions is closest to the truth.

The idea that myopic Bayesian agents will not experiment and so need not learn the truth has been explored in a number of economic models. For example, a monopolist facing an unknown demand curve might choose to set each period's price to maximize current expected profit, and then never learn what the best price would be, as in McLennan (1984). Similarly, a player in a game who never experiments with some actions might not learn how his opponents would respond to them, and a system of such myopic learners could converge to a self-confirming equilibrium whose outcome is not Nash (Fudenberg and Levine (1993b,a), Fudenberg and Kreps (1995)). The optimal "active learning" rules for correctly specified Bayesians (meaning that the true state is in the support of their prior) have also been extensively studied, notably in the multi-armed bandits of Gittins (1979) and Whittle (1980) where the payoff to each arm is independent of the payoffs of the others, the optimal stopping problems (e.g. Arrow *et al.* (1949), Chernoff (1972), Moscarini and Smith (2001), and Fudenberg *et al.* (2015)), and in learning models where the results of one action can provide information a parameter that also determines the expected returns to other actions (Easley and Kiefer (1988); Kiefer and Nyarko (1989) and Aghion *et al.* (1991)). In all of these cases, if the agent has a sufficiently low cost of information (i.e. is sufficiently patient and/or faces low flow costs of acquiring signals) she will learn enough to play the full-information optimal action. Similarly, patient rational learners with non-doctrinaire priors over opponents' strategies cannot converge to non-Nash outcomes in games (Fudenberg and Levine (1993b)); their non-doctrinaire prior assumption makes sure that each player's subjective model is correctly specified.

In all of the models with a correctly specified prior, the agent's beliefs eventually converge. With misspecified beliefs this is not the case, as shown by Dubins and Freedman (1966). Berk (1966) is the seminal paper on the asymptotic behavior of the posterior distribution when the class of models the agent considers possible does not contain the truth. Here the signals are exogenous and exchangeable, and the agent is trying to learn an unknown parameter θ. Berk (1966) showed that the posterior concentrates a.s. (with respect to the true distribution) on the subset Θ_p of the parameter set Θ on which the Kullback-Leibler divergence of the true distribution with respect to the subjective distributions is minimal. Thus, even though the posterior distributions need not converge, the support of the posterior does converge.[1]

In the econometrics literature, the maximum likelihood estimator of θ in case of misspecified econometric model is known as quasi-maximum likelihood estimator (QMLE). In the case when the Kullback-Leibler divergence is minimized at a unique belief $\Theta_p = \{\theta_p\}$, QMLE is a natural estimator of θ_p. The signal process is assumed to be exogenous, and in addition it is typically assumed that the process of observations satisfies near epoch dependence (Gallant and White (1988)), which, roughly, requires that dependence on past realizations fades away sufficiently quickly. In this case, QMLE converges to θ_p p-a.s. (see e.g., Gallant and White (1988), Theorem 3.19). In contrast to these econometrics papers, it is natural for the signal process to have long memory when there is active learning. For instance in a multiple-armed bandit problem, the very first signal realization can determine whether the agent sticks to the safe arm or continues with the risky arm forever. Here lies the key difference between our work and literature on misspecification in statistics.

It is especially natural to consider misspecification in the contest of parametric learning models, since any parametric prior (such as the assumption of a linear demand curve with unknown slope and intercept) assigns probability zero to "most" payoff functions (Nyarko (1991)). Arrow and Green (1973) discussed a number of forms of misspecification of demand in oligopoly, including linear demand with exponentially distributed parameters, and worked

[1] Shalizi (2009) gives a more general treatment of the problem: observations follow some stochastic process and the considered set of models is not parametric. He requires that for any model θ and associated density of observations f_θ, the limit of $\frac{1}{t} \log \frac{f_\theta(y_1,\dots y_t)}{p(y_1,\dots y_t)}$ exists p almost-surely, where $\{y_i\}$ are observations and p is the true density. This assumption can not be guaranteed independently of the agent's actions and is often violated in misspecified learning with endogenous signals, for example in the cycles in Nyarko (1991).

out the learning dynamics in that case; where beliefs and actions converge to a point that is determined by the priors. In contrast, beliefs and actions cycle with the two-point priors in Nyarko (1991)'s otherwise identical oligopoly model. Recently there has also been interest in misspecification as a result of behavioral biases. The initial work in this area (the cursed equilibrium of Eyster and Rabin (2005), and the analogy-based equilibrium of Jehiel (2005), Jehiel and Koessler (2008)) incorporated the misspecification directly into the definition of the equilibrium concept; Esponda (2008) provide a learning-theoretic foundation for these concepts in the case of a purely myopic learner who never experiments. Esponda and Pouzo (2015b) develop the notion of "Berk-Nash equilibrium," and show that it corresponds to a limit point of learning by myopic agents whose payoffs are perturbed by random shocks and who also make asymptotically vanishing optimization errors. , Heidhues *et al.* (2015) consider a model of learning with misspecified beliefs where the optimal active learning rule is myopic.

In our setting, the agent's action and associated signal distribution are not fixed but change endogenously over time. As far as we know, the only prior studies of learning by misspecified agents in this situation are Esponda and Pouzo (2015b,a); Heidhues *et al.* (2015); our paper differs in considering non-myopic agents and in allowing some actions to generate uninformative signals.[2]Throughout the paper we assume that the agents are doctrinaire Bayesians. It would be interesting to allow for agents who are aware of the possibility of misspecification and conduct tests to detect it, such as the stationarity tests used in Fudenberg and Kreps (1993), Sargent (1999) and Cho and Kasa (2014). We adopt a continuous-time model to avoid oscillations that otherwise occur near stationary points of the system and to obtain sharper results.

2 The Model

Time is continuous and denoted by $t \in [0, +\infty)$. At every point in time t the agent takes an action $a_t \in A$, where the set of possible actions A is finite. At time t the agent receives flow payoff $\mathrm{d}\pi_t$ and observes it. Objectively the flow payoff at time t when the agent takes the action a_t is given by,

$$\mathrm{d}\pi_t = \tilde{\pi}(a_t)\mathrm{d}t + \sigma(a_t)\mathrm{d}W_t, \tag{2.1}$$

[2] In the learning model of Esponda and Pouzo (2015b,a), payoff perturbations lead the agent to always assign positive probability to every informative action, which rules out mistaken convergence to the safe arm in a bandit problem. Also, they impose an "identifiability" assumption that requires that the models that are closest to the true model are indistinguishable given the data available to the agent. We do not assume this.

where W_t is a standard Brownian motion, and $\sigma(a) > 0$ is the volatility when the agent takes action a. The agent thinks that the only possible states of the world are $\Theta = \{0, 1\}$. In each state $i \in \{0, 1\}$ the agent believes that the flow payoff is given by

$$\mathrm{d}\pi_t = \pi^i(a_t)\mathrm{d}t + \sigma(a_t)\mathrm{d}W_t\,. \tag{2.2}$$

Note that the function σ is the same for both states and is objectively correct. We denote by a^i the action that maximizes the flow payoff in state i, which we assume is unique

$$a^i = \arg\max_{a\in A} \pi^i(a)\,.$$

We assume that there are no *informationally equivalent actions:* there is no pair of distinct actions a',a'' such that

$$\frac{\pi^1(a') - \pi^0(a')}{\sigma(a')} = \frac{\pi^1(a'') - \pi^0(a'')}{\sigma(a'')}\,.$$

This implies in particular that $a^0 \neq a^1$; without loss of generality we assume that $a^0 < a^1$.

The agent's filtration, which is generated by observation of the payoff process (π_t), is denoted by $\{\mathcal{F}_t\}_{t\geq 0}$. The set of agent's strategies, i.e. processes adapted to $\{\mathcal{F}_t\}$, is denoted by S. We use $\mathbb{P}^s[\cdot], \mathbb{E}^s[\cdot]$ to denote the agent's subjective probability measure and expectation operator when he uses strategy $s \in S$, and $\tilde{\mathbb{P}}^s[\cdot], \tilde{\mathbb{E}}^s[\cdot]$ to denote the probability measure and expectation operator of an outside observer who knows the true payoff $\tilde{\pi}$ function, and thus know the objective probability measure when the agent uses strategy s. When a strategy s has been fixed, we will write a_t for the agent's action at time t along the course of a particular realization of the process. The subjective probability the agent assigns to state 1 at time t is denoted by

$$p_t \triangleq \mathbb{P}^s[\theta = 1 \mid \mathcal{F}_t]\,.$$

$p_0 \in (0, 1)$ is the prior probability the agent assigns to state 1 at time zero. Since Θ is a doubleton,we will refer to p_t as the agent's belief.

The agent's objective is to pick a strategy to maximize the expected discounted flow of payoffs discount with rate r,

$$\max_{s\in S} \mathbb{E}^s\left[r\int_0^{+\infty} e^{-r\,t}\mathrm{d}\pi_t\right].$$

We allow for both the purely myopic case $r = \infty$ and the case of an agent with $r < \infty$ who values future payoffs and so has an incentive to experiment.

3 Illustrative Example: Seller with Unknown Linear Demand

Imagine an Amazon seller of a differentiated product. Suppose that he thinks that the elasticity of demand is constant, but is not sure how elastic it is. He already tried some initial price, and now he is considering to whether change his price. If the demand is highly elastic, it's optimal to decrease the price; if the demand is inelastic, it's optimal to increase the price. In actuality, the elasticity of demand is not constant but low for low prices and high for high prices. What is the dynamics of the seller's actions and beliefs, and what does the seller do in the long run?

Nyarko (1991) studied this problem when the seller is myopic and has only two actions: high price and low price. A main finding in Nyarko (1991) is that beliefs of a seller with a misspecified demand model need not to converge. Indeed, when the seller tries the low price, he detects low elasticity. He extrapolates this low elasticity to the entire demand curve and sets the high price. After that he detects high elasticity, which makes him set the low price, etc. Here prices and beliefs oscillate and do not converge because the distribution of price signals under the action that is optimal in state 1 is closer to the distribution the seller expects under state 0, and conversely the myopically optimal action in state 0 generates signals that suggest that increase the seller's belief in state 1.

However, the results of Nyarko (1991) depend critically on the assumptions that the player has only two actions and is completely myopic. We show that adding a third uninformative action which is myopically optimal for intermediate beliefs will lead a not too patient seller to eventually play the uninformative action forever, so that his beliefs converge. In contrast, the beliefs of a sufficiently patient seller do not converge, because he will never choose the uninformative action. The key here is that, unlike in the case of a correctly specified payoff function, a misspecified agent can continue to believe he has a non-trivial "option value" from using actions that are not myopically optimal, even in the limit as his data set grows large. The example here is the simplest way to show this qualitative distinction between the patience and myopia when players are misspecified. The rest of the paper studies a general setting with arbitrary many actions and general payoff functions.

The seller optimizes against a linear demand function and can pick among three prices, one of them uninformative:

$$A = \{-1, 0, 1\}.$$

We normalize prices and profits to simplify the algebra. The seller's perceived linear demand

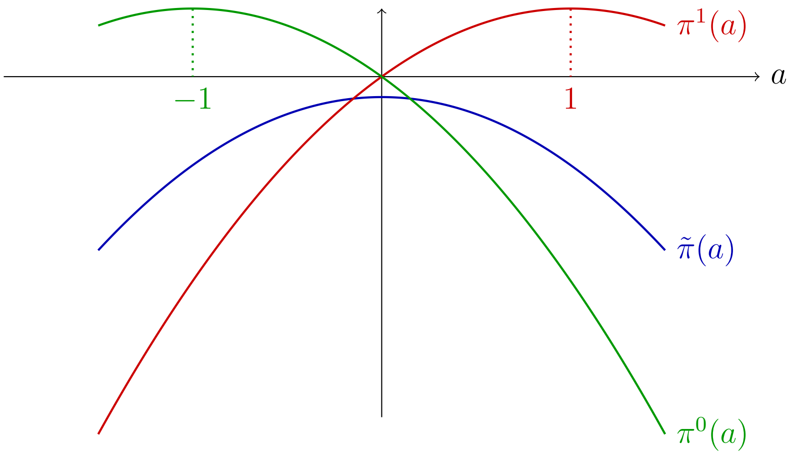

Figure 3.1: Objective and subjective payoffs.

function gives rise to quadratic subjective profit functions:

$$\begin{aligned}\pi^1(a) &= -a(a-2),\\ \pi^0(a) &= -a(a+2).\end{aligned}$$

The true profit function is, however,

$$\tilde{\pi}(a) = -a^2 - .3.$$

The signal volatility is constant in action, $\sigma(a) \equiv \sigma$. See Figure 3.1 for visualization of payoff functions.

Imagine that the agent is very confident that the state is 1 (p is close to 1) and therefore plays the positive action $a = 1$. When $a = 1$ is played, the true payoff is closer to the state-0 average payoff than to the state-1 average payoff:

$$|\pi^1(1) - \tilde{\pi}(1)| - |\tilde{\pi}(1) - \pi^0(1)| = 2 + .3 - (-.3 + 2) > 0.$$

As we will show later, this implies that state 0 is closer to the truth in the sense of Kullback-Leibler divergence, and the agent's belief drifts down and he becomes less confident in state 1. Conversely (and symmetrically), when the agent is confident that the state is state 0 (p close to 0),he plays the negative action $a = -1$, and the state-1 average payoff is closer to the truth than state-0 average payoff:

$$|\pi^1(1) - \tilde{\pi}(1)| - |\tilde{\pi}(1) - \pi^0(1)| = 2 - .3 - (.3 + 2) < 0.$$

Here the agent's belief drifts upwards and he becomes more convinced of state 1.

Therefore, this dynamics pushes agent's belief p_t from boundaries to the center of $[0, 1]$. The question is whether the agent continue the oscillation between actions 1 and -1 indefinitely, or if he will eventually play the uninformative intermediate action $a = 0$? It turns out that the answer depends on the agent's discount rate. Specifically, using the results from the main part of the paper, we will be able to show the following.

Claim 1. When the agent uses his optimal strategy, there exists a critical discount factor $\hat{r} = 6/\sigma^2$ such that:

i) When the seller is impatient $r > \hat{r}$, his actions converge almost surely to the uninformative action $\lim_{t\to\infty} a_t = 0$, and bis beliefs converge almost surely to 1/2.

ii) When the seller is patient $r < \hat{r}$, his actions and beliefs almost surely do not converge.

The patient agent's beliefs and actions do not converge because he believes that there are sufficiently large gains from learning when his belief is close to 1/2 such that it is optimal to experiment with either $a = 1$ or $a = -1$ and never take the uninformative action $a = 0$. As the optimal action under state 0 generates signals in favor of state 1, and conversely, the optimal action under state 1 generates signals in favor of state 0 the patient agent experiments indefinitely.[3]

4 The Dynamics of Optimization and Learning

4.1 Dynamics of Beliefs

We start by characterizing the evolution of the agent's beliefs with respect to his subjective probability measure. To do so we define the *informativeness* $I \colon A \to \mathbb{R}$ of an action a as

$$I(a) \triangleq \frac{\pi^1(a) - \pi^0(a)}{\sigma(a)} \,.$$

Intuitively, if $I(a) > 0$ the agent who takes action a at time t interprets higher flow payoffs as evidence of the state being θ_1, while if $I(a) < 0$ she takes high flow payoffs as evidence

[3] One might object that the player would notice that he is not converging and reconsider his model. We have two responses to this. First, in our setting any signal path realizes with positive probability, so the cycles while a priori unlikely do not flatly contradict the agent's subjective model and need not lead a Bayesian to reject it. Second, one may think of our analysis as a prediction about what happens before the (non-Bayesian) agent runs a falsification test and rejects his current model. Cho and Kasa (2014) develop this idea.

of the state being θ_0. The bigger the absolute value of $I(a)$, the more strongly the agent's belief reacts to her flow payoffs. To simplify notation define

$$\pi^p(a) \triangleq p\pi^1(a) + (1-p)\pi^0(a)$$

as the flow payoff the agent expects when taking action a when holding the belief p. For a given strategy s we define a process which measures how much the realized payoffs deviated from the agent's expected payoffs

$$Z_t^s \triangleq \int_0^t \frac{\mathrm{d}\pi_\tau}{\sigma(s_\tau)} - \int_0^t \frac{\pi_{p_\tau}(s_\tau)}{\sigma(s_\tau)}\mathrm{d}\tau\,.$$

As is well known (see Bolton and Harris (1999), Liptser and Shiryaev (1974, Theorem 9.1)), under the agent's subjective probability measure, the process Z is a Brownian motion. Furthermore, the belief $(p_t)_{t\in\mathbb{R}_+}$ is a martingale and can be characterized as a solution to the SDE[4]

$$\mathrm{d}p_t^s = p_t^s(1-p_t^s)I(a_t)\mathrm{d}Z_t^s. \tag{4.1}$$

To simplify notation we subsequently drop the explicit dependence on the strategy and denote the belief process by just p. The log belief ratio of the subjective probability of state θ_1 and state θ_0 is denoted by

$$R_t \triangleq \log\frac{p_t}{1-p_t}. \tag{4.2}$$

This transformation of the belief process to the associated log-likelihood process will be convenient for our future results.

The next lemma derives the evolution of the log likelihood under the objective probability measure. Under the objective measure, neither p nor R is a martingale, but the evolution of the log-likelihood follows from Ito's Lemma.

Lemma 1. *Given a strategy s the dynamics of the agent's belief R_t are given by*

$$\mathrm{d}R_t = I(a_t)\left[\frac{\tilde{\pi}(a_t) - \pi^{1/2}(a_t)}{\sigma(a_t)}\mathrm{d}t + \mathrm{d}W_t\right], \tag{4.3}$$

where W_t is a standard Brownian motion under the objective distribution.

[4] To avoid technicalities we assume that the agent is restricted to strategies such that Eq. (4.1) admits a unique strong solution, i.e. the posterior belief is pathwise well defined. A simple sufficient condition is the restriction to Markov strategies where the agent's action is piecewise constant in his belief.

Proof. The dynamics of the belief process are given by

$$\mathrm{d}p_t = p_t(1-p_t)I(a_t)\left[\frac{\tilde{\pi}(a_t)-\pi^{p_t}(a_t)}{\sigma(a_t)}\mathrm{d}t - \mathrm{d}W_t\right] ,$$

where W is the Brownian motion which determines the true payoff process. As $R(p) = \log\frac{p}{1-p}$ is twice differentiable we can apply Ito's Lemma and get

$$\begin{aligned}
\mathrm{d}R_t &= R'(p_t)\mathrm{d}p_t + R''(p_t)\frac{[p_t(1-p_t)I(a_t)]^2}{2}\mathrm{d}t \\
&= \frac{\mathrm{d}p_t}{p_t(1-p_t)} + \frac{1-2p_t}{[p_t(1-p_t)]^2}\frac{[p_t(1-p_t)I(a_t)]^2}{2}\mathrm{d}t \\
&= I(a_t)\left[\frac{\tilde{\pi}(a_t)-(p_t\pi^1(a_t)+(1-p_t)\pi^0(a_t))+(\frac{1}{2}-p_t)(\pi^1(a_t)-\pi^0(a_t))}{\sigma(a_t)}\mathrm{d}t - \mathrm{d}W_t\right] \\
&= I(a_t)\left[\frac{\tilde{\pi}(a_t)-(\pi^1(a_t)+\pi^0(a_t))/2}{\sigma(a_t)}\mathrm{d}t + \mathrm{d}W_t\right] . \quad \square
\end{aligned}$$

4.2 Relation to Kullback-Leibler Divergence

Let $KL(j,a)$ for $j \in \{0,1\}$ be *the Kullback-Leibler divergence* between the payoff distribution under the true state and the payoff distribution under state j, when the agent plays action $a \in A$:

$$KL(j,a) \triangleq \int_{\mathbb{R}} \log\left[\frac{\phi(x;\pi^j(a),\sigma(a))}{\phi(x;\tilde{\pi}(a),\sigma(a))}\right]\phi(x;\tilde{\pi}(a),\sigma(a))\mathrm{d}x,$$

where $\phi(x;\pi,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\pi)^2}{2\,\sigma}\right)$ is density of the normal distribution with mean π and variance σ^2. Simple algebra shows that

$$KL(j,a) = \frac{(\pi^j(a)-\tilde{\pi}(a))^2}{2\sigma^2(a)}.$$

Define $\Delta(a)$ as the difference between these two divergences when action a is played:

$$\Delta(a) \triangleq KL(0,a) - KL(1,a). \tag{4.4}$$

Note that $\Delta(a)$ is finite because $\sigma^2(a) > 0$ for all $a \in A$. In discrete time with a fixed signal generating process beliefs converge towards the subjective state whose signal distribution minimizes the Kullback-Leibler divergence to the true state (Berk (1966)). Thus, in discrete time with a fixed action a, p_t would converge to 1 if $\Delta(a) > 0$ and to 0 if $\Delta(a) < 0$. As the next result shows, $\Delta(a_t)$ determines the drift of the log-likelihood ratio (and thus the belief) process in our continuous model. As we will see below in Proposition 1, this naturally extends the discrete time result to non-constant actions.

Fact 1. *The drift of log-likelihood ratio process R given in (4.3) is equal to $\Delta(a_t)$. Its squared volatility is equal to $I(a_t)^2$, which is also the Kullback-Leibler divergence between the payoff distributions in state 1 and state 0 when the agent plays action a_t.*

This fact lets us give some intuition for Eq. (4.3). If for some $a \in A$, $\tilde{\pi}(a) > \pi^{.5}(a)$ and $I(a) > 0$, the true expected flow payoff $\tilde{\pi}(a)$ is closer to $\pi(a)$, the expected payoff in the state θ_1, than to $\pi^0(a)$, the payoff in the state θ_0. This implies that observed signals are on average closer to state 1 than state 0, or $\Delta(a) > 0$. As a result, the belief process drifts upwards.

4.3 Optimal Behavior

The next technical preliminary is to verify that an optimal policy for the agent exists. This has not yet been shown in our setting, even in the case of a correctly specified agent. The closest results are those of Strulovici and Szydlowski (2014), but their results are not immediately applicable here because they assume the variance of the controlled process is uniformly bounded from below. In our setup this assumption corresponds to assuming that the informativeness of each action is nonzero ($I(a) \neq 0$) for all $a \in A$. To circumvent this problem we recast our model as a combined optimal control and optimal stopping problem.

We call an action $a \in A$ *uninformative* if $I(a) = 0$ and denote by $U \subset A$ the set of uninformative actions $U \triangleq \{a \in A \colon I(a) = 0\}$. Note, that whenever the agent takes an uninformative action her beliefs stay constant by Eq. (4.1). As his beliefs stay constant under any uninformative action it is optimal to take the action which maximizes the expected flow payoff given his subjective belief p. As $\pi^1(a) = \pi^0(a)$ for every uninformative action a it follows that the payoff from an uninformative action is independent of the agent's belief. Hence, it is without loss of generality to assume that there is at most one uninformative action $a^u \in \operatorname{argmax}_{a \in U} \pi^0(a)$; we denote the payoff of this action by $g \triangleq \pi^0(a^u)$.

Define the value function as the highest average expected value which can be achieved by the agent using an arbitrary strategy given his initial belief p,

$$v_r(p) \triangleq \begin{cases} \sup_{s \in S} \mathbb{E}^s_p \left[r \int_0^{+\infty} e^{-rt} \mathrm{d}\pi_t \right] & \text{for } r < \infty \\ \max_a \pi^p(a) & \text{for } r = \infty \end{cases}.$$

The following theorem characterizes the value function and shows the existence of a Markovian optimal strategy.

Theorem 1. *For each interest rate $r \in (0, +\infty]$, there exists an optimal Markovian strategy $s_r^* \colon [0,1] \to A$. The value function $v_r \colon [0,1] \to \mathbb{R}_+$ is continuous, convex in the belief, and continuous in the discount rate r, with $\lim_{r\to\infty} v_r \equiv v_\infty$. Furthermore, for $r < \infty$ it is twice continuously differentiable on $\{p \in (0,1) \colon v_r(p) > g\}$ and satisfies:*

$$v_r(p) = \max_{a\in A} \pi^p(a) + [I(a)p(1-p)]^2 \, \frac{v_r''(p)}{2\,r}.$$

Any optimal strategy maximizes this expression at each $p \in (0,1)$.

The proof of the theorem, which is given in the appendix, first solves the auxiliary problem where the agent is restricted to informative actions, and once the belief leaves an exogenously specified set takes the optimal uninformative action forever. Then, the proof verifies that if this set is chosen appropriately the resulting policy from the auxiliary problem is optimal in the original problem.

The next result uses Theorem 1 to say more about the form of the optimal strategy.

Lemma 2. *The optimal strategy s_r^* has the following properties:*

i) For any r, there exists a unique interval $[\underline{u}, \overline{u}] \subset [0,1]$ such that the uninformative action is optimal if and only if $p \in [\underline{u}, \overline{u}]$.

ii) If there are no informationally equivalent actions, the optimal action $s_r^*(p)$ is unique for almost every belief $p \in [0,1]$, and the evolution of the agent's beliefs is independent of which optimal strategy he uses.

iii) There exists an interval of beliefs around $p = 0$ and $p = 1$ such that the unique optimal action is myopic, i.e. for any r, $\exists \lim_{p\searrow 0} s_r^*(p) = a^0, \lim_{p\nearrow 0} s_r^*(p) = a^1$.

In the usual setting of a correctly specified agent, one can reduce a set of informationally equivalent actions to its equivalence class without loss of generality. With a misspecified agent, though, this is not the case, and the next example shows how the conclusion of Lemma 2 can fail when there are actions that are subjectively but not objectively equivalent.

Example 1. $A = \{a^0, a', a'', a^1\}$, $\sigma \equiv 1$, and the payoff functions are as follows.

	a^0	a'	a''	a^1
$\pi^0(a)$	5	4	4	0
$\pi^1(a)$	0	1	1	5
$\tilde{\pi}(a)$	4	3	2	1

Actions a' and a'' are informationally equivalent because $I(a') = I(a'')$. For a myopic agent with some intermediate $\hat{p}$, both a' and a'' are optimal. However, the objective payoff to a' and a'' are distinct, and the learning dynamics depends on the optimal action selection. Action a' generates signals that point to state 0 while action a'' generates signals that point to state 1. Indeed, the drift of R is $\Delta(a') = -1.5 < 0$ when a' is played, and is $\Delta(a'') = 1.5 > 0$ when a'' is played.

In what follows we maintain the assumption of no informationally equivalent actions.

5 Asymptotic Beliefs and Actions: Complete Characterization

The main result of this section is that the asymptotic behavior of actions and beliefs is pinned down by by the local properties of payoff functions near the steady states, and in particular whether these steady states are absorbing or repelling.

Fix an optimal strategy s_r^*. The belief $\hat{p}$ is a *steady state* if whenever $p_t = \hat{p}$ and $a_t = s_r^*(p_t)$, we have $dp_t = 0$. By inspecting formula (4.1), one can see that first, there are two *corner steady states* $p = 0$ and $p = 1$, and second, there can be *interior steady states* at beliefs $\hat{p} \in (0,1)$ such that the optimal action $s_r^*(\hat{p})$ is uninformative. If there is an interior steady state, the process $\{p_t\}$ takes values in a strict subset of $[0,1]$ because the process has continuous sample paths and cannot pass through the steady state. Thus, the range of $\{p_t\}$ depends on the prior p_0. From Lemma 2, there is an interval of beliefs $[\underline{u}, \overline{u}]$ such that the uninformative action is optimal. Let $(\underline{p}, \overline{p})$ be the largest interval on which the subjective beliefs evolve before hitting a steady state. If p_0 is already a steady state, i.e. $p_0 \in \{0,1\} \cup [\underline{u}, \overline{u}]$, then the analysis is trivial. In what follows we study the non-trivial case. We have:

$$(\underline{p}, \overline{p}) \quad = \quad \begin{cases} (0,1) & \text{if } [\underline{u}, \overline{u}] = \varnothing \\ (\overline{u}, 1) & \text{if } 1 > p_0 > \overline{u} \\ (0, \underline{u}) & \text{if } 0 < p_0 < \underline{u} \end{cases} . \tag{5.1}$$

The range of $\{p_t\}$ is therefore $[\underline{p}, \overline{p}]$.

Setting up the model in continuous time drastically simplifies the analysis, because in discrete time the belief process can jump over steady states and oscillate near them, which makes it more complicated to define the appropriate notion of convergence and state limit results.

We call $\overline{p}$ *attracting* if there is positive objective probability that beliefs converge to $\overline{p}$, that is if $\tilde{\mathbb{P}}[\lim_{t\to\infty} p_t = \overline{p}] > 0$. We say that $\overline{p}$ is *repelling* if the objective probability of converging to $\overline{p}$ is 0. The lower bound belief $\underline{p}$ is classified the same way. The difference in Kullback-Leibler divergences $\Delta(a)$ determines which of the states, 0 or 1, is subjectively closer to the truth when the agent plays the action a. By Lemma 2 the action a^0 is played for beliefs close to zero and the action a^1 for beliefs close to one. Intuitively, the local dynamics of the belief process around $p = 0$ and $p = 1$ are hence completely determined by $\Delta(a^0)$ and $\Delta(a^1)$. The next result shows that also the long-run dynamics of the agent's belief process are completely determined by $\Delta(a^0)$ and $\Delta(a^1)$. Recall that $\Delta(a)$ is given by

$$\Delta(a) = \frac{(\pi^1(a) - \pi^0(a))(\tilde{\pi}(a) - \pi^{1/2}(a))}{\sigma^2(a)}. \tag{5.2}$$

Proposition 1. *1. $\overline{p}$ is attracting if $\overline{p} < 1$, or if $\overline{p} = 1$ and $\Delta(a^1) > 0$. $\overline{p}$ is repelling if $\overline{p} = 1$ and $\Delta(a^1) < 0$.*

2. $\underline{p}$ is attracting if $\underline{p} > 0$, or if $\underline{p} = 0$ and $\Delta(a^0) < 0$. $\underline{p}$ is repelling if $\underline{p} = 0$ and $\Delta(a^0) > 0$.

3. If both $\underline{p}$ and $\overline{p}$ are repelling, then the agent's belief objectively converges with probability zero. If either $\underline{p}$ or $\overline{p}$ is attracting, then the agent's belief objectively converges with probability one to a point in $\{\underline{p}, \overline{p}\}$.

The proof is in the appendix; here we give the intuition behind the result. For concreteness consider the upper steady state $\overline{p}$. If it is interior ($\overline{p} < 1$), the volatility of belief around it does not vanish, and there is a positive chance of hitting $\overline{p}$ even if the drift leads away from $\overline{p}$. If $\overline{p} = 1$, then the volatility vanishes as p_t approaches $\overline{p}$, and it turns out that that the sign of the drift is sufficient to determine the convergence property: positive drift makes p_t converge to $\overline{p}$ with positive probability, while negative drift prevents p_t from converging to $\overline{p}$.

Another intuition involves KL divergence. Consider $\overline{p} = 1$. Action a^1 is the last informative action before the agent hits the steady state. If at a^1, state 1 is closer to the truth than state 0 in terms of KL divergence ($\Delta(a^1) > 0$), then the agent receives the signals that on average favor state 1. This makes him willing to keep on playing a^1, and therefore he hits the upper steady state $\overline{p}$ with positive probability.

We are able to obtain this clean characterization of the beliefs asymptotics because diffusion processes in continuous time admit a sharp characterization of the limit distribution (Lemma 4 in the Appendix).

The next result shows that the discrete-time result of Berk (1966) extends to the continuous time.

Corollary 1. *Assume the agent has only one action, $A = \{a\}$. Then p_t asymptotically concentrates on the set of θ's closest to the truth in terms of KL divergence.*

Proof. There are three cases for $\Delta(a)$: negative, zero and positive. Positive (negative) $\Delta(a)$ means that $\theta = 1$ ($\theta = 0$) is closer to the truth, and $p_t \to 1$ ($p_t \to 0$) by Proposition 1. If $\Delta(a) = 0$, then both states are equally close to the truth, and the statement of the result is obvious. □

The next result shows that under the correct specification and no uninformative actions, the beliefs converge to the true state.

Corollary 2. *Assume $\tilde{\pi} = \pi^1$, and $I(a) \neq 0$ for all $a \in A$. Then $\tilde{\mathbb{P}}[p_t \to 1] = 1$.*

Proof. Since there are no uninformative actions, $(\underline{p}, \overline{p}) = (0, 1)$. Then $\Delta(a) > 0$ for all $a \in A$. Therefore, $\Delta(a^0), \Delta(a^1) > 0$. By Proposition 1, $p_t \to 1$ $\tilde{\mathbb{P}}$-a.s. □

The next proposition shows that a sufficiently patient agent does not play uninformative actions in "non-trivial" cases, where a trivial case is the one where all informative actions are dominated by uninformative actions. The proof is deferred to the Appendix.

Proposition 2. *Let there be a belief $\hat{p} \in (0, 1)$ such that all myopic best responses to $\hat{p}$ are informative. Then, for each $p \in (0, 1)$, there is $\bar{r}$ such that for $r < \bar{r}$, uninformative actions are not optimal. If additionally a^0 and a^1 are informative, then there is a uniform $\bar{r}$ such that for all $p \in [0, 1]$ only informative actions are optimal.*

Setting the model in the continuous time allows us to give a concise proof of Proposition 1. Indeed, when agent's beliefs follow a diffusion process, the posterior belief reaches any nonempty interval of $[0, 1]$ with positive probability. This implies that the gains from learning are always positive, so a very patient agent will not take an uninformative action.

A version of the following result is known in the literature and serves here as another illustration of Propositions 1 and 2: Under the correct specification, when the agent is patient enough, the limit action is optimal.

Corollary 3. *Assume that a^0 and a^1 are informative, and the model is not misspecified, i.e. $\tilde{\pi} \in \{\pi^1, \pi^2\}$. Then there exists $\bar{r} > 0$ such that for all $r < \bar{r}$ the agent's action objectively converges to the optimal action $\lim_{t\to\infty} a_t = \max_a \tilde{\pi}(a)$ with probability one.*

Proof. Without loss of generality let $\tilde{\pi} = \pi^1$. Case 1: For all $p \in [0,1]$, there is uninformative optimal action. Pick a weakly dominant optimal action $\hat{a}$. It is optimal for all p, and so trivially, the limit action is optimal. Case 2: there is a belief $\hat{p} \in (0,1)$ such that all myopic best responses to $\hat{p}$ are informative. By Proposition 2, there is a uniform $\bar{r}$ such that for all $p \in [0,1]$ only informative actions are optimal. Therefore, $(\underline{p}, \overline{p}) = (0,1)$. Now, from correct specification we have $\Delta(a^0), \Delta(a^1) > 0$, and so by Proposition 1, $\tilde{\mathbb{P}}[\lim_{t\to\infty} p_t = 1] = 1$. This implies $\tilde{\mathbb{P}}[\lim_{t\to\infty} a_t \to a^1] = 1$. □

The next proposition is the converse of Proposition 2 and states that if an uninformative action is strictly myopically optimal at some belief, then it is still optimal for slightly patient agent.

Proposition 3. *Suppose there is a $\hat{p} \in [0,1]$ such that an uninformative action is the myopically strict best response to $\hat{p}$. Then, there is $\underline{r}$ such that for $r > \underline{r}$, an uninformative action is a best response to $\hat{p}$.*

Proof. Let S_r be the maximal set of beliefs where an uninformative action is optimal: $S_r := \{p \in [0,1] \colon v_r(p) = g\}$. We have that $\hat{p} \in \mathit{int}\, S_{+\infty}$. By Theorem 1, v is continuous in r, consequently S_r is continuous in r. Therefore, for $\underline{r}$ large enough, $\hat{p} \in S_r$ for all $r > \underline{r}$. □

6 Examples

6.1 Seller with Unknown Linear Demand

In this section we prove the results we stated in Claim 1. Recall

$$\begin{aligned} A &= \{-1, 0, 1\}, \\ \pi^1(a) &= -a(a-2), \\ \pi^0(a) &= -a(a+2), \\ \tilde{\pi}(a) &= -a^2 - \eta, \quad \eta > 0, \\ \sigma(a) &\equiv \sigma. \end{aligned} \tag{6.1}$$

The proof is an illustration of how to apply the general results of Proposition 1, 2 and 3 to a particular situation. The analysis follows 2 steps: characterize the range of beliefs $[\underline{p}, \overline{p}]$ using Proposition 2, and then apply Proposition 1.

Proof of Claim 1. Use Lemma 1 to find

$$\mathrm{d}R_t = -4a_t \cdot \eta\sigma^{-2} dt + 4\sigma^{-1} a_t \mathrm{d}W_t, \tag{6.2}$$

where W_t is the objective noise process. The drift of R is equal to $\Delta(a) = -4a\eta/\sigma^2$, and also $I(a) = 4a\sigma^{-1}$.

First we show that the myopic agent's beliefs and actions converge to the interior steady state where he plays the uninformative action $a = 0$. We start by computing the value function of the myopic agent:

$$\begin{aligned} v_\infty(p) &= \max_a \pi^p(a) = \max\{\pi^p(1), \pi^p(-1), \pi^p(0)\} \\ &= \max\{p + (1-p)(-3), -3p + 1 - p, 0\} \\ &= \max\{4p - 3, -4p + 1, 0\}. \end{aligned}$$

EveryThe optimal myopic strategy satisfies

$$s^*_\infty(p) = \begin{cases} 1, & \text{for } p > 3/4 \\ 0, & \text{for } p \in (1/4, 3/4) \\ -1, & \text{for } p < 1/4 \end{cases}.$$

If $p_0 \in [1/4, 3/4]$, the agent plays the uninformative action so the system is in the steady state. If $p_0 > 3/4$, then $\underline{p} = 3/4$, $\overline{p} = 1$. We have

$$\Delta(1) = -\frac{4\eta}{\sigma^2} < 0. \tag{6.3}$$

Apply Proposition 1 and find that $\overline{p}$ is repelling, and $\underline{p}$ is attracting. Therefore, p_t converges to $\underline{p} = 3/4$ almost surely. The case of $p_0 < 1/4$ is completely symmetric. Here, p_t converges to $1/4$ almost surely. As a result, whatever the initial belief , the beliefs converge to the interior steady state almost surely if the agent is myopic.

Second, we show that beliefs and actions of a sufficiently patient agent do not converge. By Proposition 2, there is $\bar{r}$ such that there are no interior steady states. In this case, $\underline{p} = 0$ and $\overline{p} = 1$. We have that $\Delta(a^1) < 0$ and $\Delta(a^0) > 0$. Applying Proposition 1 we see that $\overline{p}$ and $\underline{p}$ are both repelling and thus the objective probability that the agent's beliefs converge equals zero by Proposition 1.

Lastly, we find the critical value $\hat{r} = 6/\sigma^2$ by solving the differential equation for the value function. By Theorem 1, the value function is characterized by HJB equation

$$v_r(p) = \max_{a \in A} \pi^p(a) + [I(a)p(1-p)]^2 \frac{v''(p)}{2r}.$$

If $r < \hat{r}$, we have to have that the maximum in the above expression is attained on $A \backslash U =$

$\{-1, 1\}$. By symmetry, action 1 is optimal on $p > 1/2$ and -1 is optimal on $p < 1/2$. Therefore, on $p \in (1/2, 1)$, the differential equation for v is

$$v_r(p) = \pi^p(1) + [I(1)p(1-p)]^2 \frac{v_r''(p)}{2\,r} = -3 + 4p + \left(\frac{4}{\sigma}\right)^2 p^2(1-p)^2 \frac{v_r''(p)}{2\,r}.$$

This differential equation admits a closed form solution

$$v_r(p) = -3 + 4p + (1-p)^\beta p^{1-\beta} C_1 + (1-p)^{1-\beta} p^\beta C_2,$$

where C_1, C_2 are free constants, and

$$\beta = \frac{1}{2} + \frac{1}{2}\sqrt{\frac{4+\alpha}{\alpha}}, \qquad \alpha = \left(\frac{4}{\sigma}\right)^2 \frac{1}{2r}.$$

The differentiability of v and symmetry implies that $v_r'(1/2) = 0$. Also, $v_r(1) = 1$. From these initial conditions we find $C_2 = 0$ and $C_1 = 4/(2\beta - 1)$. Therefore,

$$v(p) = -3 + 4p + \frac{4}{2\beta - 1}(1-p)^\beta p^{1-\beta}. \tag{6.4}$$

Since $a = 0$ is never played when $r < \hat{r}$, it has to be the case that

$$v(1/2) \geq 0.$$

The equality is satisfied when

$$r \leq \frac{6}{\sigma^2}.$$

□

6.2 Amazon Seller with Two Prices

Here we modify the previous examples by supposing that there are only two feasible actions, both of which are informative.

$$A = \{-1, 1\}.$$

This is the case of Nyarko (1991): 1 is the high price, -1 is the low price. Note that the uninformative action $a = 0$ is unavailable. It is immediate that beliefs do not converge

Claim 2. $\tilde{\mathbb{P}}[p_t \text{ converges}] = \tilde{\mathbb{P}}[a_t \text{ converges}] = 0$.

Proof. There are no interior steady states by the choice of action space. Both $\underline{p}$ and $\overline{p}$ are repelling as shown above. Therefore, the system does not converge by Proposition 1. □

Intuitively, at $a = 1$, the belief drift is negative, and at $a = -1$, the belief drift is positive. We get that even the myopic agent cannot get to $a = 0$ because it is unavailable! So we get cycles a la Nyarko. To sum up, Nyarko's cycles have different nature and are independent of patience/impatience. It is not to his favor because it is unreasonable to assume that a monopolist can set only one of two prices. If he assumed continuous set of prices, he would get convergence. The similarity between his myopic agent and our patient agent is that in both cases the action space is effectively restricted to not contain the uninformative action $a = 0$. But it Nyarko's myopic case the restriction is exogenous, while in our patient case $a = 0$ is not played by endogenous reasons.

6.3 A Bandit Model of Learning

Suppose now that there are two actions a^1, a^0, the second one of which we call safe and assume to be uninformative

$$\pi^0(a^0) = \pi^1(a^0) = s \in \mathbb{R}_+ \,.$$

We call the other action risky and assume that it leads to a high expected payoff $h \in \mathbb{R}_+$ in state θ_1 of the world and a low payoff $l \in \mathbb{R}_+$ otherwise

$$h = \pi^1(a^1) > \pi^0(a^1) = l \,.$$

It is easy to see that the optimal strategy of the agent is to take the risky-action if and only if his posterior likelihood is above a threshold $R^\star$. We write $\sigma = \sigma(a^1)$ for the noise level when the risky arm is chosen. the threshold $R^\star$ depends on s, l, h, σ as well as the discount factor r. let us denote by π the true expected payoff of the risky arm

$$\pi = \tilde{\pi}(a^1) \,.$$

The true payoff of the safe arm is completely irrelevant for the agents behavior in this example as the agent will stick with the safe arm forever once he has chosen the safe arm for the first time. By Lemma 1, the dynamics of the posterior likelihood of the agent R_t are given by

$$\mathrm{d}R_t = \mathbf{1}_{\{R_t > R^\star\}} \frac{(h-l)}{\sigma} \left[\frac{\pi - \frac{h+l}{2}}{\sigma} \mathrm{d}t + \mathrm{d}W_t \right] \,.$$

Thus, the posterior likelihood is a Brownian motion with drift $(h-l)\frac{\pi - \frac{h+l}{2}}{\sigma^2}$ which is absorbed at $R^\star$.

We consider only the non-trivial case when $R_0 > R^*$. We have that $\overline{R} = +\infty$, $\underline{R} = R^*$.

From (5.2) we find that

$$\Delta(a^0) = \Delta(a^1) = \frac{(h-l)}{\sigma} \frac{\pi - \frac{h+l}{2}}{\sigma}.$$

By Proposition 1, $\underline{R}$ is attracting because it is finite. Next, $R = +\infty$ is attracting if $\pi > \frac{h+l}{2}$, and repelling if $\pi < \frac{h+l}{2}$. Thus we have proved the following result:

Claim 3. If the true payoff of the risky arm π is closer to l than h, then the agent eventually switches to the safe arm with probability 1. If the true payoff of the risky arm is closer to h than l, then with some probability strictly positive probability the agent sticks to the risky arm forever.

7 Conclusion

This paper has given a first look at active learning by a misspecified Bayesian agent. As we have seen, even with only two subjectively possible states the dynamics depend on the agent's interest rate, as well as on the availability of an uninformative action. Still as in traditional two-state Bayesian models, the agent's subjective probability over states is a sufficient condition for his beliefs, so the speed of belief updating is constant no matter how much data the agent has already obtained.

References

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

Arrow, K. and Green, J. (1973) Notes on Expectations Equilibria in Bayesian Settings, mimeo.

Arrow, K. J., Blackwell, D. and Girshick, M. A. (1949) Bayes and minimax solutions of sequential decision problems, *Econometrica, Journal of the Econometric Society*, pp. 213–244.

Berk, R. (1966) Limiting Behavior of Posterior Distributions when the Model is incorrect, *The Annals of Mathematical Statistics.*

Bolton, P. and Harris, C. (1999) Strategic experimentation, *Econometrica*, **67**, 349–374.

Chernoff, H. (1972) *Sequential analysis and optimal design*, vol. 8, Siam.

Cho, I.-K. and Kasa, K. (2014) Learning and model validation, *The Review of Economic Studies*.

Diaconis, P. and Freedman, D. (1986) On the Consistency of Bayes Estimates, *The Annals of Statistics*, **14**, 1–26.

Dubins, L. and Freedman, D. (1966) Invariant probabilities for certain Markov processes, *The Annals of Mathematical Statistics*, **222**.

Easley, D. and Kiefer, N. (1988) Controlling a Stochastic Process with Unknown Parameters, *Econometrica*, **56**, 1045–1064.

Esponda, I. (2008) Behavioral Equilibrium in Economies with Adverse Selection, *The American Economic Review*, **98**, 1269–1291.

Esponda, I. and Pouzo, D. (2015a) A Framework for Modeling Bounded Rationality: Misspecified Bayesian-Markov Decision Processes, working paper.

Esponda, I. and Pouzo, D. (2015b) Berk-Nash Equilibrium: A Framework for Modeling Agents with Misspecified Models, working paper.

Eyster, E. and Rabin, M. (2005) Cursed equilibrium, *Econometrica*, **73**, 1623–1672.

Fudenberg, D. and Kreps, D. (1993) Learning Mixed Equilibria, *Games and Economic Behavior*.

Fudenberg, D. and Kreps, D. M. (1995) Learning in extensive-form games I. Self-confirming equilibria, *Games and Economic Behavior*, pp. 20–55.

Fudenberg, D. and Levine, D. (1993a) Self-Confirming Equilibrium, *Econometrica*, **61**, 523–545.

Fudenberg, D. and Levine, D. (1993b) Steady State Learning and Nash Equilibrium, *Econometrica*, **61**, 547–573.

Fudenberg, D., Strack, P. and Strzalecki, T. (2015) Stochastic choice and optimal sequential sampling, working paper.

Gallant, A. R. and White, H. (1988) *A unified theory of estimation and inference for nonlinear dynamic models*, Basil Blackwell New York.

Gittins, J. C. (1979) Bandit processes and dynamic allocation indices, *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 148–177.

Heidhues, P., Koszegi, B. and Strack, P. (2015) Unrealistic expectations and misguided learning, *Available at SSRN.*

Jehiel, P. (2005) Analogy-Based Expectation Equilibrium, *Journal of Economic theory*, pp. 1–38.

Jehiel, P. and Koessler, F. (2008) Revisiting games of incomplete information with analogy-based expectations, *Games and Economic Behavior*, **62**, 533–557.

Karatzas, I. and Shreve, S. (2012) *Brownian motion and stochastic calculus*, vol. 113, Springer Science & Business Media.

Kiefer, N. and Nyarko, Y. (1989) Optimal Control of an Unknown Linear Process with Learning, *International Economic Review*, **30**, 571–586.

Liptser, R. and Shiryaev, A. N. (1974) *Statistics of Random Processes*, Nauka.

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

Moscarini, G. and Smith, L. (2001) The optimal level of experimentation, *Econometrica*, **69**, 1629–1644.

Nyarko, Y. (1991) Learning in mis-specified models and the possibility of cycles, *Journal of Economic Theory*, **55**, 416–427.

Sargent, T. J. (1999) The conquest of american inflation, *Princeton, NJ: Princeton.*

Shalizi, C. R. (2009) Dynamics of Bayesian updating with dependent data and misspecified models, *Electronic Journal of Statistics*, **3**, 1039–1074.

Strulovici, B. H. and Szydlowski, M. (2014) On the smoothness of value functions and the existence of optimal strategies, working paper.

Whittle, P. (1980) Multi-armed bandits and the gittins index, *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 143–149.

Appendix

A Proofs

A.1 Proofs of Theorem 1 and Lemma 2

For any closed set let τ_D be the first hitting time of $D \subseteq [0,1]$

$$\tau_D = \inf\{t \geq 0 \colon p_t \in D\}.$$

The following auxiliary result will be useful to establish the existence of an optimal policy. It shows that an optimal Markovian policy exists when the agent is restricted to informative strategies and he switches to the optimal uninformative action once his belief reaches an exogenously given set D. The maximal average payoff from playing the uninformative action forever equals

$$g \triangleq \pi^0(a^u)\,. \tag{A.1}$$

The proof follows the Strulovici and Szydlowski (2014) analysis of a two-state, multi armed bandit where all arms are informative, but uses the log-likelihood as the state instead of the probability of state 1.[5]

Lemma 3. *For any closed set $D \subseteq [0,1]$ the control problem*

$$\max_a r\mathbb{E}\left[\int_0^{\tau_D} e^{-r\,\tau_D}\pi^p(a)\mathrm{d}t + e^{-r\tau_D}g\right]$$

where the agent is restricted to informative controls $a_t \in A \backslash U$ admits a value function which is twice differentiable on $[0,1] \setminus D$ and solves

$$v(p) = \max_{a \in A\backslash U} \pi^p(a) + [I(a)p(1-p)]^2\,\frac{v''(p)}{2\,r}.$$

Any optimal policy maximizes this expression at each p.

Proof of Lemma 3. We will use Strulovici and Szydlowski (2014) to establish the result. As the variance of the belief process is not uniformly bounded from below the conditions of Strulovici and Szydlowski (2014) are not satisfied directly (their result needs the diffusion coefficient $I(a)p(1-p)$ to be uniformly bounded away from zero). To avoid this problem we

[5] If we use the belief p as a state variable, the diffusion coefficient $I(a)p(1-p)$ is not uniformly bounded away from zero, so the conditions for the existence of a classical solution are not satisfied.

will use the log-likelihood ratio process R_t instead of the belief process p_t as a state variable

$$R_t \triangleq \log \frac{p_t}{1-p_t}$$

We denote by $w(R) \triangleq v(p(R))$ the value function of the agent when he holds the belief

$$p(R) \triangleq \frac{e^R}{e^R+1} .$$

For further reference let us note that

$$\begin{aligned} p'(R) &= \frac{e^R}{(1+e^R)^2} = (1-p(R))p(R) && \text{(A.2)} \\ p''(R) &= p'(R)(1-2p(R)) . && \text{(A.3)} \end{aligned}$$

The dynamics of (R_t) are given by

$$\mathrm{d}R_t = \left[p(R_t) - \frac{1}{2}\right] I(a_t)^2 \mathrm{d}t + I(a_t)\mathrm{d}W_t ,$$

where (W_t) is a Brownian motion according to the agent's subjective probability measure. Note, that the drift term $\mu(R,a) \triangleq \left[p(R_t) - \frac{1}{2}\right] I(a)_t^2$ is bounded by

$$|\mu(R,a)| \leq \frac{1}{2} \max_{a'} I(a')^2 ,$$

the variance is bounded from above and below by

$$\min_{a' \in A \setminus U} I(a')\sigma(a') \leq I(a)\sigma(a) \leq \max_{a' \in A \setminus U} I(a')\sigma(a')$$

and the flow payoffs are bounded by

$$|\pi_{p(R)}(a)| \leq \max_{a} \max\{|\pi^1(a)|, |\pi^0(a)|\} .$$

Furthermore, the variance term is independent of R and thus Lipschitz continuous, the flow payoff and the drift of R are linear in the belief p and as the belief is Lipschitz continuous in R they are Lipschitz continuous in R as well. Thus, Assumption 1, 2 and 3 from Strulovici and Szydlowski (2014) are satisfied and it follows that the value function $w : \mathbb{R} \to \mathbb{R}$ is twice differentiable and satisfies

$$w(R) = \max_{a \in A \setminus U} \pi_{p(R)}(a) + \frac{I(a)^2}{r} \left(\left[p(R) - \frac{1}{2}\right] w'(R) + \frac{1}{2} w''(R) \right) . \tag{A.4}$$

Note, that as $w(R) = v(p(R))$ and by Eq. A.2 and A.3 we have that

$$\begin{aligned}
w'(R) &= v'(p(R))p'(R) = v'(p)\,p(1-p) \\
w''(R) &= v''(p(R))[p'(R)]^2 - p''(R)v'(R) \\
&= v''(p)[p(1-p)]^2 - 2[p - \frac{1}{2}]w'(R)\,.
\end{aligned}$$

Plugging into Eq. A.4 yields

$$v(p) = \max_{a \in A \setminus U} \pi^p + I(a)^2 \frac{[p(1-p)]^2}{2\,r} v''(p)\,. \qquad \square$$

Proof of Theorem 1. First, note that v is well defined as an upper bound on v is the payoff from taking the optimal action forever

$$v(p) \leq p\left[\max_a \pi^1(a)\right] + (1-p)\left[\max_a \pi^0(a)\right]\,,$$

and a lower bound is given by taking the action which is optimal for the belief p forever

$$v(p) \geq \max_a \pi^p(a)\,.$$

As v is the supremum over linear function v is convex. As every convex function is continuous v is continuous and the set of beliefs $\mathcal{S} \subseteq [0,1]$ for which an uninformative action is optimal is closed

$$\mathcal{S} \triangleq \{p \in [0,1] \colon g = v(p)\}\,.$$

As $\mathcal{S}$ is closed we can define the first time the belief process reaches $\mathcal{S}$, $\tau_{\mathcal{S}} \triangleq \inf\{t \colon p_t \in \mathcal{S}\}$. Define $\hat{x}$ as the process which is absorbed in $\mathcal{S}$

$$\hat{x}_t = x_{\min\{t, \tau_{\mathcal{S}}\}}\,.$$

Consider the control problem where the agent is restricted to informative controls $a_t \in A \setminus U$ and the process is absorbed the first time it leaves $[0,1]\setminus\mathcal{S}$ with a payoff of $g(p)$. By Lemma 3 the value function $\hat{v} : [0,1] \setminus \mathcal{S} \to \mathbb{R}$ of this problem is twice differentiable, and solves

$$\hat{v}(p) = \max_{a \in A \setminus U} \pi^p(a) + [I(a)p(1-p)]^2 \, \frac{\hat{v}''(p)}{2\,r}\,. \tag{A.5}$$

with boundary condition $\hat{v}(p) = g$ for all p on the boundary of $[0,1] \setminus \mathcal{S}$. Furthermore, an optimal Markovian control $a^\star : [0,1] \setminus \mathcal{S} \to A \setminus U$ exists.

We extend this policy into a Markovian policy on $[0,1]$ by setting

$$a^\star(p) = \arg\max_{a\in U} \pi^p(a)\,,$$

for all $p \in \mathcal{S}$.

We first prove that the value function satisfies $v(p) = \hat{v}(p)$ for all $p \in [0,1] \setminus \mathcal{S}$. To prove this we show that it is never optimal to use an uninformative action in $[0,1] \setminus \mathcal{S}$ and thus $\hat{v}(p) = v(p)$ by the definition of $\hat{v}$. Suppose, it is optimal to chose an uninformative action at the belief $p \in [0,1] \setminus \mathcal{S}$ for a (random) time $\hat{\tau}$. Then, as the belief does not change prior to $\hat{\tau}$

$$\begin{aligned}
v(p) &= r\mathbb{E}\left[\int_0^{\hat{\tau}} e^{-r\,t}\pi^{p_t}(a_t)\mathrm{d}t + \int_{\hat{\tau}}^{\infty} e^{-r\,t}\pi^{p_t}(a_t)\mathrm{d}t \mid p_0 = p\right] \\
&= (1-\mathbb{E}\left[e^{-r\hat{\tau}}\right])g(p) + \mathbb{E}\left[e^{-r\hat{\tau}}\right]v(p) \\
\Rightarrow 0 &= (1-\mathbb{E}\left[e^{-r\hat{\tau}}\right])\left(g(p)-v(p)\right)\,.
\end{aligned}$$

As $g < v(p)$ for all $p \in [0,1] \setminus \mathcal{S}$ by definition of $\mathcal{S}$ it follows that $\hat{\tau} = 0$. Thus, it is never optimal to chose an uninformative action for a positive amount of time on $p \in [0,1] \setminus \mathcal{S}$ and $\hat{v}(p) = v(p)$.

Finally, we verify that $a^\star$ is an optimal policy: The verification for $p \in \mathcal{S}$ follows immediately from the definition of $\mathcal{S}$. The verification argument for $p \in [0,1] \setminus \mathcal{S}$ is standard and uses the fact that the value function is twice differentiable on $[0,1] \setminus \mathcal{S}$ to apply Ito's Lemma. Fix an arbitrary policy a using the law of iterated expectations and the definition of the stopping set $\mathcal{S}$ yields

$$\begin{aligned}
r\mathbb{E}\left[\int_0^{\infty} e^{-r\,t}\pi^{p}(a_t)\mathrm{d}t\right] &= r\mathbb{E}\left[\int_0^{\tau_{\mathcal{S}}} e^{-r\,t}\pi^{p_t}(a_t)\mathrm{d}t + \mathbb{E}\left[\int_{\tau_{\mathcal{S}}}^{\infty} e^{-r\,t}\pi^{p_t}(a_t)\mathrm{d}t \mid p_{\tau_{\mathcal{S}}}\right]\right] \\
&\le r\mathbb{E}\left[\int_0^{\tau_{\mathcal{S}}} e^{-r\,t}\pi^{p_t}(a_t)\mathrm{d}t + e^{-r\,\tau_{\mathcal{S}}}g\right]\,. \qquad \text{(A.6)}
\end{aligned}$$

In the next step we use that by Eq. A.5 for all $p \in [0,1] \setminus \mathcal{S}$ and all $a \notin U$ we have

$$\pi^p(a) \le \hat{v}(p) - [I(a)p(1-p)]^2\,\frac{\hat{v}''(p)}{2\,r}\,.$$

For $a \in U$ we have that by definition of $\mathcal{S}$ and the fact that $\hat{v}(p) = v(p)$ for $p \in [0,1] \setminus \mathcal{S}$

$$\pi^p(a) < v(p) = \hat{v}(p) - [I(a)p(1-p)]^2\,\frac{\hat{v}''(p)}{2\,r}\,.$$

By Ito's Lemma and Doob's optional sampling Theorem we have that

$$\begin{aligned}\mathbb{E}\left[\int_0^{\tau_S} r\,e^{-r\,t}\pi^{p_t}(a_t)\mathrm{d}t\right] &\leq \mathbb{E}\left[\int_0^{\tau_S}\left\{r\,e^{-r\,t}\hat{v}(p)-e^{-r\,t}\left[I(a)p(1-p)\right]^2\frac{\hat{v}''(p)}{2}\right\}\mathrm{d}t\right]\\ &= \mathbb{E}\left[\hat{v}(p_0)-e^{-r\,p_{\tau_S}}\hat{v}(p_{\tau_S})\right]\,. \qquad (A.7)\end{aligned}$$

Combining Eq. A.6 and Eq. A.7 and the fact that $\hat{v}(p_{\tau_S})=g$ by the boundary condition of Eq. A.5 yields,

$$r\mathbb{E}\left[\int_0^{\infty} e^{-r\,t}\pi^p(a_t)\mathrm{d}t\right]\leq\hat{v}(p_0)=r\mathbb{E}\left[\int_0^{\infty} e^{-r\,t}\pi^p(a^\star(p_t))\mathrm{d}t\right]\,.$$

Thus, for any policy a the value is lower than the value when following the policy $a^\star$.

In the last step we verify that the value function is Lipschitz continuous in r. The derivative of the value with respect to the interest rate for a fixed strategy equals

$$\begin{aligned}\left|\mathbb{E}\left[\int_0^{\infty}(1-r\,t)\,e^{-r\,t}\pi^p(a_t)\mathrm{d}t\right]\right| &\leq \mathbb{E}\left[\int_0^{\infty}(1+r\,t)\,e^{-r\,t}\left|\pi^p(a_t)\right|\mathrm{d}t\right]\\ &\leq \mathbb{E}\left[\int_0^{\infty}(1+r\,t)\,e^{-r\,t}\mathrm{d}t\right]\max_{\theta\in\{0,1\}}\max_a\left|\pi_\theta(a)\right|\\ &= \tfrac{2}{r}\max_{\theta\in\{0,1\}}\max_a\left|\pi_\theta(a)\right|\,.\end{aligned}$$

It thus follows from the envelope Theorem that the value function is Lipschitz continuous in r for all r bounded away from zero. To see that v is continuous in r at $r=0$, observe that an upper bound on the agent's payoff is given by the payoff the agent obtains when knowing the state and taking the optimal action

$$v_r(p)\leq p\pi^0(a^0)+(1-p)\pi^1(a^1)\,. \qquad (A.8)$$

The agent can take an informative action for a long, but deterministic, time T to learn the state arbitrarily precisely and afterwards take an optimal action. As the agent becomes patient his loss in payoff from the initial experimentation phase vanishes and thus $\liminf_{r\to 0} v_r(p)$ exists and equals the payoff the agent could obtain when knowing the state. As the payoff the agent could obtain when knowing the state is also an upper bound, the limit exists and we have

$$\lim_{r\to 0} v_r(p)=p\pi^0(a^0)+(1-p)\pi^1(a^1)\equiv v_0(p)\,.$$

Finally, we argue that the agent value function converges for $r\to\infty$. First, note that as the agent can always play the myopic optimum given his initial belief, and ignore all subsequent information we have

$$\liminf_{r\to\infty} v_r(p) \geq \max_a \pi^p(a) = v_\infty(p)\,.$$

To see that $v_\infty(p)$ is also an upper bound, observe that the agent's payoff can not be better than if he used the full-information optimal action after time $\tau > 0$

$$\begin{aligned} v_r(p) &\leq \mathbb{E}\left[\int_0^\tau r\,e^{-r\,t}\pi^{p_t}(a_t)\mathrm{d}t + e^{-r\tau}v_0(p_\tau) \mid p_0 = p\right] \\ &\leq (1-e^{-r\tau})\mathbb{E}\left[\sup_{t\in[0,\tau]} \pi^{p_t}(a_t) \mid p_0 = p\right] + e^{-r\tau}\mathbb{E}\left[v_0(p_\tau) \mid p_0 = p\right]\,. \end{aligned}$$

As $\pi^{p_t}(a_t) \leq \max_a \pi^{p_t}(a) = v_\infty(p_t)$, we have

$$v_r(p) \leq (1-e^{-r\tau})\mathbb{E}\left[\sup_{t\in[0,\tau]} v_\infty(p_t)) \mid p_0 = p\right] + e^{-r\tau}\mathbb{E}\left[v_0(p_\tau) \mid p_0 = p\right]\,.$$

Choose $\tau = 1/\sqrt{r}$

$$v_r(p) \leq (1-e^{-\sqrt{r}})\mathbb{E}\left[\sup_{t\in[0,\tau]} v_\infty(p_t)) \mid p_0 = p\right] + e^{-\sqrt{r}}\mathbb{E}\left[v_0(p_\tau) \mid p_0 = p\right]\,.$$

Then in the limit $r \to \infty$ the second term in this sum vanishes, and we have

$$\begin{aligned} \lim_{r\to\infty} v_r(p) &\leq \lim_{r\to\infty} \mathbb{E}\left[\sup_{t\in[0,1/\sqrt{r}]} \pi^{p_t}(a_t) \mid p_0 = p\right] \\ &\leq \lim_{r\to\infty} \mathbb{E}\left[\sup_{t\in[0,1/\sqrt{r}]} v_\infty(p_t) \mid p_0 = p\right]. \end{aligned}$$

As $v_\infty(p)$ is continuous in p and almost every realization of the belief process $(p_t)_t$ is continuous in t for any strategy, we have that $\lim_{r\to\infty}\sup_{t\in[0,1/\sqrt{r}]} v_\infty(p_t) = v_\infty(p)$. And because for almost every path of the belief process v_∞ is bounded, the dominated convergence theorem implies that

$$v_\infty(p) = \lim_{r\to\infty} \mathbb{E}\left[\sup_{t\in[0,1/\sqrt{r}]} v_\infty(p_t) \mid p_0 = p\right] = \mathbb{E}\left[\lim_{r\to\infty} \sup_{t\in[0,1/\sqrt{r}]} v_\infty(p_t) \mid p_0 = p\right] = v_\infty(p)$$

so $\lim_{r\to\infty} v_r(p) \leq v_\infty(p)\,.$ □

Proof of Lemma 2. We first argue that any optimal strategy can only use the uninformative action when beliefs are in a (possibly empty) interval $[\underline{u}, \overline{u}]$. Note that at a belief p where the uninformative action is optimal, beliefs and actions will not change in the future, so at such p we have $v_r(p) = g$. Now suppose to the contrary that the set of beliefs where the uninformative action is optimal is disconnected. This implies that there are beliefs $p < p' < p''$ such that $g = v_r(p) = v_r(p'') \neq v_r(p')$. As the agent can always take the uninformative action forever, $v_r(p') \geq g$. But convexity of v_r implies that $v_r(p') \leq g$ and hence $v(p') = g$.

Next we prove that the optimal action is unique for almost every belief. If not, then since there are only finitely many actions there exists an interval $[p', p'']$ such that at least two actions $a, a' \in A \setminus U$ are optimal for any belief $p \in [p', p'']$. As both actions are optimal in $[p', p'']$ the value function solves simultaneously the two linear second order ODE's on $[p', p'']$

$$\begin{aligned} v_r(p) &= \pi^p(a) + [I(a)p(1-p)]^2 \frac{v_r''(p)}{2\,r} \\ v_r(p) &= \pi^p(a') + [I(a')p(1-p)]^2 \frac{v_r''(p)}{2\,r}\,. \end{aligned}$$

Take the difference between these ODE's to find

$$0 = \pi^p(a) - \pi^p(a') + [p(1-p)]^2 \frac{v_r''(p)}{2\,r}(I(a)^2 - I(a')^2)$$

The right hand side is not equal to zero uniformly on $[p', p'']$ from the assumption that there are no informationally equivalent actions, $I(a) \neq I(a')$. Therefore these ODE's have different solutions, which contradicts that both actions are optimal in $[p', p'']$.

Finally, as the informative action is almost everywhere unique, and the solution to the SDE of the belief process 4.3 remains unchanged when switching informative actions on a set of beliefs of Lebesgue measure zero, it follows that the beliefs of the agent are independent of which optimal strategy she uses.

To show the third part, suppose to the contrary that in any neighborhood of $p = 1$, there is a belief such that some $\hat{a} \neq a^1$ is optimal. Denote the points of indifference between a^1 and $\hat{a}$ by $\{p^k\}$. We have that $p^k \to 1$ as $k \to \infty$.

By Theorem 1, $v''(p)$ exists on $(0,1)\backslash\{v(p) > g\}$, and because the agent is indifferent between a^1 and $\hat{a}$ at each p^k the following equalities hold:

$$\begin{aligned} v(p^k) &= \pi^{p^k}(a^1) + (I(a^1)p^k(1-p^k))^2 \frac{v''(p^k)}{2r} \\ v(p^k) &= \pi^{p^k}(a^1) + (I(a^1)p^k(1-p^k))^2 \frac{v''(p^k)}{2r} \end{aligned}$$

We know that $\lim_{p\to 1} v(p) = \pi^1(a^1)$, so the first equality implies that

$$\lim_{k\to\infty} (I(a^1)p^k(1-p^k))^2 \frac{v''(p^k)}{2r} = 0$$

However, the second equality implies that

$$\lim_{k\to\infty} (I(\hat{a})p^k(1-p^k))^2 \frac{v''(p^k)}{2r} = \pi^1(a^1) - \pi^1(\hat{a}) \neq 0.$$

Taking the ratio, we obtain

$$\lim_{k\to\infty} \frac{I(a^1)^2}{I(\hat{a})^2} = 0,$$

which is a contradiction. □

A.2 Proofs of Propositions 1 and 2

Let the diffusion process $\{R_t\}$ be defined on $(\underline{R}, \overline{R})$ by $dR_t = \alpha(R_t)dt + \beta(R_t)dW_t$.

Fix arbitrary $R_0 \in (\underline{R}, \overline{R})$. The natural scale function for R is (strictly increasing and invertible) function $\phi : (\underline{R}, \overline{R}) \to \mathbb{R}$ defined by

$$\phi(R) = \int_{R_0}^{R} \exp\left(- \int_{R_0}^{y} \frac{2\alpha(z)}{\beta(z)^2} \mathrm{d}z \right) \mathrm{d}y \,. \tag{A.9}$$

Lemma 4. *Let $T = \inf\{t \geq 0 \colon R_t = \overline{R}$ or $R_t = \underline{R}\}$. Every weak solution of Eq. (4.3) has the following properties*

1. *If $\phi(\underline{R}+) = -\infty$, $\phi(\overline{R}-) = \infty$, then*

$$\mathbb{P}\left[T = \infty\right] = 1\,.$$

2. *If $\phi(\underline{R}+) > -\infty$, $\phi(\overline{R}-) = \infty$, then the probability that the process is absorbed in l with probability one*

$$\mathbb{P}\left[\lim_{t\to T} R_t = \underline{R}\right] = 1.$$

3. *If $\phi(\underline{R}+) = -\infty$, $\phi(\overline{R}-) < \infty$, then the probability that the process is absorbed in $\overline{R}$ with probability one*

$$\mathbb{P}\left[\lim_{t\to T} R_t = \overline{R}\right] = 1\,.$$

4. *If $\phi(\underline{R}+) < -\infty$, $\phi(\overline{R}-) < \infty$, then the probability that the process is absorbed in $\overline{R}$*

($\underline{R}$) is given by

$$\mathbb{P}\left[\lim_{t\to T} R_t = \underline{R}\right] = 1 - \mathbb{P}\left[\lim_{t\to T} R_t = \overline{R}\right] = \frac{\phi(\overline{R}-) - \phi(R_0)}{\phi(\overline{R}-) - \phi(\underline{R}+)} .$$

Proof. Proposition 5.22 (p.345) in Karatzas and Shreve (2012). □

Lemma 5. *Let $f \colon \mathbb{R} \to \mathbb{R}$ be such that $\lim_{y\to+\infty} yf(y) = \delta$, and let*

$$I = \int_0^{+\infty} \exp\left\{-\int_z^x f(y)dy\right\} dx$$

for some $z \in \mathbb{R}$. Then $I = +\infty$ if $\delta < 1$ and $I < +\infty$ if $\delta > 1$.

Proof. First, consider the case $\delta > 1$. Pick $M > 0$ such that $yf(y) \geq \delta' > 1$ for $y > M$. Then for $x > M$:

$$\begin{aligned} \int_z^x f(y)dy &= \int_z^M f(y)dy + \int_M^x f(y)dy > K_1 + \int_M^x (\delta'/y)\, dy \\ &= K_1 + \delta' \log x, \end{aligned}$$

where K_1 is some constant. Then

$$\begin{aligned} I &= \int_0^M + \int_M^{+\infty} \exp\left\{-\int_z^x f(y)dy\right\} dx < K_2 + \int_M^{+\infty} \exp\left\{-K_1 - \delta' \log x\right\} dx \\ &= K_2 + K_3 \int_M^{+\infty} x^{-\delta'} dx, \end{aligned}$$

where K_2 and $K_3 > 0$ are some constants. Since $\delta' > 1$, $I < +\infty$.

The case $\delta < 1$ is shown in the similar way. Pick $M > 0$ such that $yf(y) \leq \delta' < 1$ for $y > M$. Then for $x > M$:

$$\int_z^x f(y)dy = \int_z^M f(y)dy + \int_M^x f(y)dy < K_1 + \int_M^x (\delta'/y)\, dy = K_1 + \delta' \log x.$$

$$\begin{aligned} I &= \int_0^M + \int_M^{+\infty} \exp\left\{-\int_z^x f(y)dy\right\} dx > K_2 + \int_M^{+\infty} \exp\left\{-K_1 - \delta' \log x\right\} dx \\ &= K_2 + K_3 \int_M^{+\infty} x^{-\delta'} dx. \end{aligned}$$

Since $\delta' < 1$, $I = +\infty$. □

Lemma 6. *Let* $f\colon (c, +\infty) \to \mathbb{R}$ *be such that* $\lim_{y\to c}(y-c)f(y) = \delta$, *and let*

$$I = \int_c^{\bar{x}} \exp\left\{-\int_z^x f(y)dy\right\} dx$$

for some $z, \bar{x} > c$. *Then* $I = +\infty$ *if* $\delta > 1$ *and* $I < +\infty$ *if* $\delta < 1$.

Proof. We first prove the lemma with $c = 0$, and then the general case is obtain by change of variables in the integrals. First, consider the case $\delta > 1$. Pick $\varepsilon > 0$ such that $yf(y) \geq \delta' > 1$ for $y \in (0, \varepsilon)$. Then for $x < \varepsilon$:

$$\begin{aligned} \int_z^x f(y)dy &= \int_z^\varepsilon f(y)dy + \int_\varepsilon^x f(y)dy < K_1 + \int_\varepsilon^x (\delta'/y)\, dy \\ &= K_1 + \delta' \log x, \end{aligned}$$

where K_1 is some constant. Then

$$\begin{aligned} I &= \int_\varepsilon^{\bar{x}} + \int_0^\varepsilon \exp\left\{-\int_z^x f(y)dy\right\} dx > K_2 + \int_0^\varepsilon \exp\left\{-K_1 - \delta' \log x\right\} dx \\ &= K_2 + K_3 \int_0^\varepsilon x^{-\delta'} dx, \end{aligned}$$

where K_2 and $K_3 > 0$ are some constants. Since $\delta' > 1$, $I = +\infty$.

The case $\delta < 1$ is shown in the similar way. Pick $\varepsilon > 0$ such that $yf(y) \leq \delta' < 1$ for $y \in (0, \varepsilon)$. Then for $x < \varepsilon$:

$$\int_z^x f(y)dy = \int_z^\varepsilon f(y)dy + \int_\varepsilon^x f(y)dy > K_1 + \int_\varepsilon^x (\delta'/y)\, dy = K_1 + \delta' \log x.$$

$$\begin{aligned} I &= \int_\varepsilon^{\bar{x}} + \int_0^\varepsilon \exp\left\{-\int_z^x f(y)dy\right\} dx < K_2 + \int_0^\varepsilon \exp\left\{-K_1 - \delta' \log x\right\} dx \\ &= K_2 + K_3 \int_0^\varepsilon x^{-\delta'} dx. \end{aligned}$$

Since $\delta' < 1$, $I < +\infty$.

For general $c \in \mathbb{R}$ do exchange of variables twice and transform the integral to the case of $c = 0$:

$$\begin{aligned} I &= \int_c^{\bar{x}} \exp\left\{-\int_z^x f(y)dy\right\} dx = \int_0^{\bar{x}-c} \exp\left\{-\int_z^{\eta+c} f(y)dy\right\} d\eta \\ &= \int_0^{\bar{x}-c} \exp\left\{-\int_{z-c}^{\eta} f(\xi + c)d\xi\right\} d\eta\,. \end{aligned}$$

We can find that $\delta = \lim_{\xi \to 0} \xi f(\xi + c) = \lim_{y \to c}(y - c) f(y)$. □

Proof of Proposition 1. We study the non-trivial case of $\underline{p} < \overline{p}$, in which p_0 is not a steady state. Change the state variable from p to $R = \log \frac{p}{1-p}$ and find $\overline{R}$ and $\underline{R}$ which correspond to $\overline{p}$ and $\underline{p}$. Fix an optimal strategy selection $s_r^* \in \mathcal{S}$. We are going to use the well-known result in the literature of diffusion processes which says that to get the limit distribution of the process it is sufficient to evaluate the natural scale function (A.9) at boundaries $\underline{R}$ and $\overline{R}$. Fix arbitrary z and consider

$$\begin{aligned}
\phi(\overline{R}) &= \int\limits_{R_0}^{\overline{R}} \exp\left\{ -\int_z^x \frac{2\alpha(y)}{\beta^2(y)} dy \right\} dx, \\
\phi(\underline{R}) &= \int\limits_{R_0}^{\underline{R}} \exp\left\{ -\int_z^x \frac{2\alpha(y)}{\beta^2(y)} dy \right\} dx,
\end{aligned}$$

where $\alpha(R) = \frac{I(a)}{\sigma(a)} \left(\tilde{\pi}(a) - \pi^{1/2}(a) \right) = \Delta(a)$ is the drift of R_t, where $a = s_r^*(R)$, and $\beta(R) = I(a)$ is the volatility of R_t, see (4.3). Therefore,

$$\frac{2\alpha(R)}{\beta^2(R)} = 2\frac{\Delta(s_r^*(R))}{I(s_r^*(R))^2}.$$

By definition of $\underline{R}$ and $\overline{R}$, $\frac{2\alpha(R)}{\beta^2(R)}$ is finite for all $R \in (\underline{R}, \overline{R})$.

We start with the upper boundary $\overline{R}$. There are two cases: when $\overline{R}$ is infinite and when it is finite. Let's first consider the limit $R \to +\infty$, which corresponds to $\overline{R} = +\infty$. We want to apply Lemma 5, and for that we need to calculate the following limit:

$$\delta \equiv \lim_{R \to +\infty} \frac{2\alpha(R)}{\beta^2(R)} R.$$

By Lemma 2, the limit $\lim_{R \to \overline{R}} s_r^*(R)$ exists and equals a^1, and we can find that

$$\delta = \lim_{R \to +\infty} 2\frac{\Delta(a^1)}{I(a^1)^2} R = \begin{cases} +\infty, & \Delta(a^1) > 0 \\ -\infty, & \Delta(a^1) < 0 \end{cases}.$$

By Lemma 5, $\phi(+\infty) = +\infty$ if $\delta < 1$, and $\phi(+\infty) < +\infty$ if $\delta > 1$. Therefore:

$$\phi(+\infty) \begin{cases} = +\infty, & \Delta(a^1) < 0 \\ < +\infty, & \Delta(a^1) > 0 \end{cases}.$$

By Lemma 4, $R = +\infty$ is attracting if $\phi(+\infty) < +\infty$ and repelling if $\phi(+\infty) = +\infty$.

Similarly, consider the limit $R \to -\infty$ ($\underline{R} = -\infty$).

$$\lim_{R\to-\infty} \frac{2\alpha(R)}{\beta^2(R)} R \quad = \lim\nolimits_{R\to-\infty} 2\tfrac{\Delta(a^0)}{I(a^0)^2} R = \begin{cases} +\infty, & \Delta(a^0) < 0 \\ -\infty, & \Delta(a^0) > 0 \end{cases}$$

Again apply Lemma 5 and find:

$$\begin{aligned} \Delta(a^0) > 0 \quad &\Rightarrow \quad \phi(-\infty) = -\infty, \\ \Delta(a^0) < 0 \quad &\Rightarrow \quad \phi(-\infty) > -\infty. \end{aligned}$$

By Lemma 4, $R = -\infty$ is attracting if $\phi(-\infty) > -\infty$ and repelling if $\phi(-\infty) = -\infty$.

Now let's do the cases when the boundary R is finite (boundary p is interior). Start with lower boundary $\underline{R}$. We want to apply Lemma 6, and for that we need to compute

$$\delta \equiv \lim_{R\to\underline{R}} \frac{2\alpha(R)}{\beta^2(R)}(R - \underline{R}) \quad = \lim\nolimits_{R\to\overline{R}} 2(R - \underline{R})\Delta(a^0) = 0 < 1.$$

By Lemma 6 , $\phi(\underline{R}) > -\infty$. By Lemma 4, $\underline{R}$ is attracting for process $\{R_t\}$. Completely analogously, we find that if $\overline{R}$ is finite, process $\{R_t\}$ gets absorbed into $\overline{R}$ with positive probability.

Finally, Lemma 4 implies that convergence has a zero-one property. Therefore part 3 of the proposition follows. □

Proof of Proposition 2. The case when all actions are informative is trivial. Let there be at least one uninformative action, and recall that g is the maximal payoff from playing an uninformative action forever. By the premise of the proposition, there is $\hat{p}$ such that $v_\infty(\hat{p}) \triangleq \max_a \pi^p(a) > g$. We need to show that for any $r < \bar{r}$ and any p, $v_r(p) > g$.

Consider the following strategy: Play an informative action $\hat{a}$ for a fixed time interval $(0, \tau)$; play the myopic best response to p_τ throughout after. The value function from following this strategy is denoted by $\tilde{V}_r(p)$. For any $p \in (0, 1)$, we have

$$\begin{aligned} \tilde{V}_r(p) - g \quad &= \quad \mathbb{E}\left[\int_0^\tau r\, e^{-r\,t} \pi^{p_t}(\hat{a})\mathrm{d}t + e^{-r\tau} v_\infty(p_\tau) \mid p_0 = p\right] - g. \\ &= \quad \mathbb{E}\left[(1 - e^{-r\tau})(\pi^{p_t}(\hat{a}) - g) + e^{-r\tau}\left(v_\infty(p_\tau) - g \mid p_0 = p\right)\right] \\ &= \quad (1 - e^{-r\tau})(\pi^p(\hat{a}) - g) + e^{-r\tau}\, \mathbb{E}\left[v_\infty(p_\tau) - g \mid p_0 = p\right] . \end{aligned}$$

By assumption, there exists $\hat{p}$, such that $v_\infty(\hat{p}) - g > 0$. By the continuity of the value function there hence also exists an interval around $\hat{p}$ such that $v_\infty(\cdot) - g > 0$ for every point in that interval. When $\hat{a}$ is played, and $p_0 \in (0,1)$, distribution of p_τ has full support on $[0,1]$, and so $\mathbb{E}\left[v_\infty(p_\tau) - g\right] > 0$. As $\mathbb{E}\left[v_\infty(p_\tau) - g \mid p_0 = p\right]$ and $\pi^{p_0}(\hat{a}) - g$ is independent of r, we have that $\tilde{V}_r(p) - g > 0$ if r is sufficiently close to zero. Since the optimal strategy cannot do worse, $v_r(p) \geq \tilde{V}_r(p) > g$.

When a^1 and a^0 are informative, by Lemma 2 for each r there are $0 < p' < p'' < 1$ such that a^1 is optimal for $p \in (p'', 1]$, and a^0 is optimal for $p \in [0, p')$. On $p \in [p', p'']$, the function $p \mapsto \mathbb{E}\left[v_\infty(p_\tau) - g \mid p_0 = p\right]$ is bounded away from zero, as it is continuous and strictly positive. Therefore there is uniform $\bar{r}$ such that $\tilde{V}(p) > g$ for all $r < \bar{r}$ and all $p \in [p', p'']$. Putting all three intervals together, we find that for all $p \in [0,1]$, only informative actions are optimal. □