Theoretical Economics 12 (2017), 1155–1189 1555-7561/20171155

Active learning with a misspecified prior

Drew Fudenberg
Department of Economics, Massachusetts Institute of Technology

Gleb Romanyuk
Department of Economics, Harvard University

Philipp Strack
Department of Economics, University of California Berkeley

We study learning and information acquisition by a Bayesian agent whose prior belief is misspecified in the sense that it assigns probability 0 to the true state of the world. At each instant, the agent takes an action and observes the corresponding payoff, which is the sum of a fixed but unknown function of the action 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 illustrates a novel interaction between misspecification and the agent's subjective discount rate.

Keywords. Active learning, learning in games, misspecified models.

JEL classification. D83, D90.

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 ignores the value of future information, and even if the agent correctly processes the information she receives, she may repeatedly take actions that would not be optimal under full information. In contrast, the optimal rule for a patient agent requires "active learning," meaning that the agent trades off the future expected gains from experimentation against its cost in terms of foregone current payoff. The details of these active learning rules have been extensively studied in the case where 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

Drew Fudenberg: Drew.Fudenberg@gmail.com
Gleb Romanyuk: glebromanyuk@g.harvard.edu
Philipp Strack: philipp.strack@gmail.com
We are grateful for comments from Jerry Green, Kevin He, Johannes Horner, Paul Heidhues, Bruno Strulovici, Sergey Vorontsov, and two anonymous referees. National Science Foundation Grant 1643517 provided financial support.

Copyright © 2017 The Authors. Theoretical Economics. The Econometric Society. Licensed under the Creative Commons Attribution-NonCommercial License 4.0. Available at http://econtheory.org.
DOI: 10.3982/TE2480

is misspecified, in the sense that it assigns probability 0 to the true state of the world, because the space of possible models is quite large (Diaconis and Freedman 1986).

This paper is the first study of active learning and information acquisition by a misspecified Bayesian.[1] In our model, at each instant, the agent takes an action and observes the corresponding payoff, which is the sum of a fixed but unknown payoff function and an additive error term whose distribution the agent knows corresponds to the increment of a Brownian motion. The agent thinks there are two possible payoff functions, yet the true payoff function is neither of them. 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 signals in favor of another steady state. Using this fact, we show that if 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 do 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 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 her 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, 1993a, 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, and in optimal stopping problems (e.g., Wald 1945, Arrow et al. 1949, Chernoff 1972, Moscarini and Smith 2001, and Fudenberg et al. 2017). Active learning has also been studied in models of parametric learning like ours, where the results of one action can provide information about a parameter that also determines the expected returns to others (e.g., 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 sufficiently low flow costs of acquiring signals), then with high probability she will learn enough to play the full-information optimal action. Similarly, patient rational learners with nondoctrinaire and correctly specified priors over opponents' strategies cannot converge to non-Nash outcomes in games (Fudenberg and Levine 1993b).

A marked difference between these models with a correctly specified prior and those with misspecified priors is that with a correctly specified prior, the agent's beliefs eventually converge, while they need not do so when the prior is misspecified

[1] We do not provide microfoundations for why the agent has a misspecified prior, but take the misspecification as given and characterize the resulting behavior.

(Dubins and Freedman 1966). Berk (1966) is the seminal paper on the asymptotic behavior of the posterior distribution when the agent is trying to passively learn a parameter from a series of exogenous and exchangeable signals but none of the parameters the agent considers possible corresponds to the true distribution. Berk (1966) shows 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, for generic priors with finite support, the beliefs converge to a single point.[2]

In the econometrics literature, the maximum likelihood estimator of θ in the case of a misspecified econometric model is known as a 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 a 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 context of parametric learning models since any parametric prior (such as the assumption of a linear demand curve with unknown slope and intercept) assigns probability 0 to "most" payoff functions (Nyarko 1991). Moreover, in many economic applications, it is natural to suppose that the agent's action and associated signal distribution are not fixed but change endogenously over time. Arrow and Green (1973) discussed a number of forms of misspecification of demand in oligopoly, including linear demand with exponentially distributed parameters, and worked out the learning dynamics for myopic agents. In contrast, beliefs and actions cycle with the two-point priors in Nyarko's (1991) otherwise identical oligopoly model. Cursed equilibrium (Eyster and Rabin 2005) and analogy-based equilibrium (Jehiel 2005, Jehiel and Koessler 2008) incorporate misspecification directly into their definitions. Esponda (2008) provides a learning-theoretic foundation for an equilibrium with misspecified beliefs in the case of a purely myopic learner who never experiments. Esponda and Pouzo (2016a) define *Berk–Nash equilibrium*, which relaxes Nash equilibrium by replacing the requirement that players' beliefs are correct with the requirement that each player's belief minimizes the Kullback–Leibler divergence to her observations on the support of her prior. They show that if payoffs are subject to small

[2] 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 $(1/t)\log(f_\theta(y_1,\dots,y_t)/p(y_1,\dots,y_t))$ exists p almost surely, where $\{y_i\}$ are observations and p is the true density. This assumption cannot 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).

random shocks and if intended play converges to a limit "stable profile," that profile must be a Berk–Nash equilibrium if players are myopic (Theorem 2) or if the game is "weakly identified" (Theorem 4).[3] Their Supplement develops the dynamics of learning in an example with myopic players, and uses stochastic approximation to show that play converges to a mixed Berk–Nash equilibrium. They also show in Theorem 3 that any weakly identified Berk–Nash equilibrium is stable for some priors if, in addition to the payoff shocks, players make asymptotically vanishing optimization errors. Esponda and Pouzo (2016b) extend the definition of Berk–Nash equilibrium to dynamic Markov decision problems where the agents need not be myopic. While Esponda and Pouzo (2016a, 2016b) characterize the stable outcomes in a very general environment, they do not analyze when there is convergence and when particular outcomes are stable, while our work completely characterizes the convergence of beliefs and actions in a specific setting. In that sense our approaches are complementary. Heidhues et al. (2017) consider a model of learning with misspecified beliefs where the optimal active learning rule is myopic so that the agent never sacrifices short-run payoff to acquire more information.

We adopt a continuous-time model with signals generated by a controlled diffusion to obtain sharper results; we briefly discuss a discrete time analog of our model after developing the continuous-time version formally. The next section lays out the primitives of our model of misspecified learning in continuous time. Section 3 provides an example in which the agent's belief converges for high discount rates but cycles if the discount rate is low. In Section 4, we describe the dynamics of belief updating and characterize the agent's optimal policy. In Section 5, we show how the agent's optimal willingness to experiment with informative actions depends on her patience level. Section 6 establishes that the asymptotic behavior of actions and beliefs is pinned down by local properties of steady states. In Section 7, we apply our techniques to the prominent examples from the literature. Section 8 concludes.

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

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

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

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

[3]The assumed payoff shocks can have very small support so need not induce experimentation; their role is to ensure that if beliefs converge, then play does too, as in Fudenberg and Kreps (1993) and the subsequent literature on smooth fictitious play. We define weak identification and relate it to our model in Remark 2 below.

Note that the function σ is the same for both states and is objectively correct. We say that the agent's prior is *misspecified* if there is no $\theta \in \Theta$ such that $\pi^\theta = \tilde{\pi}$. We denote by a^θ the action that maximizes the flow payoff in state θ, which we assume is unique:

$$a^\theta = \operatorname*{argmax}_{a \in A} \pi^\theta(a).$$

We call these the *full-certainty* actions and assume that $a^0 \neq a^1$, as otherwise the problem is trivial.[4]

Assumption 1 (No informationally equivalent actions). *There is no pair of distinct actions a', a'' such that*

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

The agent's filtration, which is generated by observation of the payoff process (π_t), is denoted by $\mathcal{F} \triangleq (\mathcal{F}_t)_{t \geq 0}$. The set of the agent's strategies, i.e., processes adapted to $\mathcal{F}$ taking values in A, is denoted by $\mathcal{S}$.[5] We use $\mathbb{P}^s[\cdot]$ and $\mathbb{E}^s[\cdot]$ to denote the agent's subjective probability measure and expectation operator when she uses strategy $s \in \mathcal{S}$, and use $\tilde{\mathbb{P}}^s[\cdot]$ and $\tilde{\mathbb{E}}^s[\cdot]$ to denote the probability measure and expectation operator of an outside observer who knows the true payoff $\tilde{\pi}$ function, and thus knows the objective probability measure when the agent uses strategy s. When a strategy s has been fixed, we 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

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

where $p_0 \in (0, 1)$ is the prior probability the agent assigns to state 1 at time zero. Since Θ is a doubleton, we refer to p_t as the agent's belief. To simplify notation, define $\pi^{(p)}(a)$ as the flow payoff the agent expects when taking action a when holding the belief p:

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

Note that $\pi^{(1)} = \pi^1$ (when the agent is sure the state is 1, her expected payoff is given by π^1) and similarly that $\pi^{(0)} = \pi^0$.

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

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

For any finite r the agent attaches some value to future payoffs and so may choose to experiment. We also consider the case of a purely myopic agent who cares only about the current payoff; her objective is simply $\max_{a \in A} \mathbb{E}[\pi^{(p)}(a)]$.

[4] If the action $a^\star$ is optimal independent of the state, it maximizes the expected flow payoff for any belief, so the optimal strategy is simply to take the action $a^\star$ after every history.

[5] As we explain in Remark 1, our results extend to the case where the agent can use mixed strategies.

3. Illustrative example: Seller with unknown linear demand

Imagine a seller of a differentiated product. Suppose that she thinks that the elasticity of demand is constant, but is not sure how elastic it is. He already tried some initial price and now she is deciding whether the price should be changed. If demand is highly elastic, it is optimal to decrease the price, while if the demand is inelastic, it would be optimal to increase the price. In actuality, the elasticity of demand is not constant but is low for low prices and high for high prices. What is the dynamic 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. The main finding in Nyarko (1991) is that the beliefs of a seller with a misspecified demand model need not converge. Indeed, when the seller tries the low price, she detects low elasticity. She extrapolates this low elasticity to the entire demand curve and sets the high price. After that, she detects high elasticity, which makes her 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 increase the probability the seller assigns to 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 that is myopically optimal for intermediate beliefs leads the seller to eventually play the uninformative action forever, so that her beliefs converge, provided that the seller is not too patient. In contrast, the beliefs of a sufficiently patient seller do not converge, because she never chooses 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 she has a nontrivial "option value" from using actions that are not myopically optimal, even in the limit as her data set grows large. The example here is the simplest way to show this qualitative distinction between patience and myopia when players are misspecified.[6]

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 function gives rise to quadratic subjective profit functions:

$$\pi^1(a) = -a(a-2),$$
$$\pi^0(a) = -a(a+2).$$

[6]Nyarko's model is different from ours in that her state space Θ is a convex subset of $\mathbb{R}^2$, but the same cycling occurs when there are two states and two actions, as when the action 0 is removed from our example here.

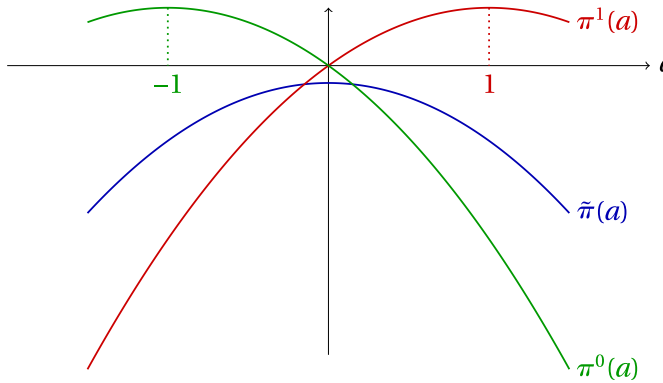

FIGURE 1. Objective and subjective payoffs.

The true profit function is, however,

$$\tilde{\pi}(a) = -a^2 - 0.3.$$

The signal volatility is constant in action, $\sigma(a) \triangleq \sigma$. See Figure 1 for a depiction of the payoff functions.

When p is close to 1, the agent plays $a = 1$. When $a = 1$ is played, the true payoff is closer to the state-0 average payoff than to the state-1 average payoff, because

$$\left|\pi^1(1) - \tilde{\pi}(1)\right| - \left|\tilde{\pi}(1) - \pi^0(1)\right| = 1 + 1.3 - (-1.3 + 3) > 0.$$

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

$$\left|\pi^1(-1) - \tilde{\pi}(-1)\right| - \left|\tilde{\pi}(-1) - \pi^0(-1)\right| = |-3 + 1.3| - |-1.3 - 1| < 0.$$

Here the agent's belief drifts upward and she becomes more convinced of state 1.

Therefore, this dynamics pushes the agent's belief p_t from the boundaries to the center of $[0, 1]$. The question is whether the agent continues to oscillate between actions 1 and -1 indefinitely or if she eventually plays the uninformative intermediate action $a = 0$, which is absorbing. 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 are be able to show the following claim.

CLAIM 1. *When the agent uses her optimal strategy, there exists a critical discount rate $\hat{r} = 6/\sigma^2$ such that the following statements hold:*

(i) When the seller is impatient, $r > \hat{r}$, her action converges almost surely to the uninformative action, $\lim_{t\to\infty} a_t = 0$, and her belief converges almost surely to $1/2$.

(ii) When the seller is patient, $r < \hat{r}$, her action and belief almost surely do not converge.

The patient agent's beliefs and actions do not converge because she believes that there are large gains from learning when her belief is close to $1/2$, so 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.[7]

Esponda and Pouzo (2016a) propose an alternative modification to the two-action Nyarko example: Instead of adding an uninformative action, they allow the agent to randomize, and define a Berk–Nash equilibrium as a point where the agent's beliefs minimize the expected KL divergence with the truth, where the expectation is over the distributions of beliefs that arises from the agent's mixed action. Our results suggest that whether play converges to this Berk–Nash equilibrium (and thus whether it is stable in the sense of Esponda and Pouzo 2016a) when the agent engages in active learning can depend on the discount factor. However, this suggestion is not definitive due to the differences between our models. In particular, we have not extended our model and results to allow for the payoff perturbations that they use to generate randomized play.

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 her subjective probability measure. To do so we define the *informativeness* $I : A \to \mathbb{R}$ of an action a as

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

Note that Assumption 1 implies that no two actions can have the same informativeness. 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 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. For a given strategy s, we define a process that measures how much the realized payoffs deviated from the agent's expected payoffs

$$Z_t^s \triangleq \int_0^t \frac{\mathrm{d}\pi_\tau}{\sigma(a_\tau)} - \int_0^t \frac{\pi_{p_\tau}(a_\tau)}{\sigma(a_\tau)}\,\mathrm{d}\tau. \tag{1}$$

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

[7] One might object that the player would notice that she is not converging and reconsider her 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 her current model. Fudenberg and Kreps (1993), Sargent (1999), and Cho and Kasa (2015) develop this idea.

to the stochastic differential equation[8] (SDE)

$$\mathrm{d}p_t^s = p_t^s\big(1 - p_t^s\big)I(a_t)\,\mathrm{d}Z_t^s. \tag{2}$$

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

$$L_t \triangleq \log\frac{p_t}{1-p_t}.$$

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 L is a martingale, but the evolution of the log likelihood L follows from applying Ito's lemma on the dynamics of p.

Lemma 1. *Given a strategy s, the dynamics of the agent's log-likelihood ratio L_t are given by*

$$\mathrm{d}L_t = I(a_t)\left[\frac{\tilde{\pi}(a_t) - \pi^{(1/2)}(a_t)}{\sigma(a_t)}\,\mathrm{d}t + \mathrm{d}W_t\right], \tag{3}$$

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

4.2 *Relation to Kullback–Leibler divergence*

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

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

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

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

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

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

[8]To avoid technicalities we assume that the agent is restricted to strategies such that (2) admits a unique strong solution, i.e., the posterior belief is pathwise well defined. A simple sufficient condition is a restriction to Markov strategies where the agent's action is piecewise constant in her belief.

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 toward 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 see below in Proposition 3, this naturally extends the discrete time result to nonconstant actions.

Fact 1. *The drift of log-likelihood ratio process L given in (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 (3). If for some $a \in A$, $\tilde{\pi}(a) > \pi^{(1/2)}(a)$, and $I(a) > 0$, then the true expected flow payoff $\tilde{\pi}(a)$ is closer to $\pi^1(a)$—the expected payoff in the state $\theta = 1$—than it is to $\pi^0(a)$—the payoff in the state $\theta = 0$. 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 upward.

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 (2015), 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 $I(a) = 0$, which is true if and only if $\pi^1(a) = \pi^0(a)$. The payoff of any uninformative action must be independent of the state, so in generic games every optimal policy will only use the single uninformative action with the highest subjective payoff. Denote this action by a^u and denote its payoff by $g \triangleq \pi^0(a^u)$.

Define the value function as the highest average expected value that can be achieved by the agent using an arbitrary strategy given her initial belief p:

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

The following theorem characterizes the value function and shows the existence of a Markovian optimal strategy, i.e., a strategy that depends on the history only through the current belief.

THEOREM 1. *For each discount rate $r \in (0, +\infty]$, there exists an optimal strategy $s_r^* : [0,1] \to A$, which is Markovian. The value function $v_r : [0,1] \to \mathbb{R}_+$ is continuous, convex in the belief p, continuous and nondecreasing in the discount rate r, and $\lim_{r\to\infty} v_r(p) = v_\infty(p)$ for all p. Furthermore, for $r < \infty$, $v_r(p)$ is twice continuously differentiable on $\{p \in (0,1) : v_r(p) > g\}$ and satisfies the Hamilton–Jacobi–Bellman (HJB) equation*

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

Any optimal strategy maximizes this expression for $p = p_t$ after almost every history.

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 fact that v is nonincreasing in r follows, as increasing r is equivalent to increasing the volatility of the signals, and the agent could replicate this by adding the noise herself.

REMARK 1 (Mixed strategies). Because the supremum over all strategies is attained by a Markov strategy, it is clear that the agent could not gain from an ex ante randomization over the space of all possible pure strategies. The same conclusion holds if instead we define mixed strategies as the limit of high-frequency oscillations among pure actions, as the resulting Bellman equation is linear in the probability of taking each action.[9]

5. ACTIVE EXPERIMENTATION

This section shows how the agent's optimal willingness to experiment with informative actions depends on her patience level.

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

LEMMA 2. *An optimal strategy s_r^* has the following properties:*

(i) For any r, there exists a unique (but possibly empty) 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) 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 she 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, $\exists \lim_{p\searrow 0} s_r^*(p) = a^0$, $\lim_{p\nearrow 1} s_r^*(p) = a^1$.

[9]Formally, we extend the payoff functions $\tilde{\pi} : \Delta(A) \to \mathbb{R}$ and the volatility $\sigma : \Delta(A) \to \mathbb{R}$ to mixed actions by taking the average over payoffs/volatilities when the agent's action is distributed according to $\beta \in \Delta(A)$, i.e., $\tilde{\pi}(\beta) = \sum_a \beta_a \tilde{\pi}(a)$ and $\sigma(\beta) = (\sum \beta_a \sigma^2(a))^{1/2}$. The informativeness $I(\beta)$ is defined as the average informativeness, $I(\beta) = \sum_a \beta_a I(a)$.

Note that because beliefs evolve continuously, if the prior belief is below $\underline{u}$, it can never rise above it, and that if the prior is above $\overline{u}$, it cannot fall below it.

The next example shows how the conclusion of Lemma 2 part (ii) can fail if we drop Assumption 1 and allow actions that are subjectively but not objectively equivalent.

Example 1. The set of feasible actions is $A = \{a^0, a', a'', a^1\}$ and the signal variance is $\sigma(a) \equiv 1$ for all a, 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 L is $\Delta(a') = -1.5 < 0$ when a' is played and is $\Delta(a'') = 1.5 > 0$ when a'' is played. ◊

Our first main result shows that a sufficiently patient agent does not play uninformative actions.

Proposition 1. *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 $r < \bar{r}$ and all $p \in [0,1]$ only informative actions are optimal.*

The proof is deferred to the Appendix; it constructs a strategy that shows that the gains from learning outweigh the loss from experimentation when the agent is sufficiently patient.

The next proposition is the partial converse of Proposition 1 and states that if an uninformative action is strictly myopically optimal at some belief, then it is still optimal when the agent is slightly patient.

Proposition 2. *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: $S_r \triangleq \{p \in [0,1] : v_r(p) = g\}$. We have that $\hat{p} \in \operatorname{int} S_{+\infty}$. 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. Asymptotic beliefs and actions: Complete characterization

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

Fix an optimal strategy s_r^*. The belief $\hat{p}$ is a *steady state belief* if whenever $p_t = \hat{p}$ and $a_t = s_r^*(p_t)$, we have $\mathrm{d}p_t = 0$. By inspecting formula (2), 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 $\hat{p} \in (0, 1)$ where the optimal action $s_r^*(\hat{p})$ is uninformative. Action $\hat{a}$ is a *steady state action* if there is a steady state belief $\hat{p}$ such that $\hat{a} = s_r^*(\hat{p})$. The two full-certainty actions a^0 and a^1 are steady states. If there is an interior steady state belief, the corresponding steady state action must be a^u.

We say that action $a \in A$ is *attracting* if there is positive objective probability that action converges to a, that is, if $\tilde{\mathbb{P}}[\lim_{t\to\infty} a_t = a] > 0$. We say that $a \in A$ is *repelling* if the objective probability of converging to a is 0.

The next result completely classifies steady state actions into attracting/repelling and shows that the long-run dynamics of the belief process is completely determined by the properties of payoff functions evaluated at steady state actions. The interior steady state is always attracting. Whether a corner steady state is attracting or repelling is determined by the sign of the difference in Kullback–Leibler divergences $\Delta(a)$ evaluated at the full-certainty action. It is convenient to rewrite $\Delta(a)$ as given by

$$\Delta(a) = \frac{\big(\pi^1(a) - \pi^0(a)\big)\big(\tilde{\pi}(a) - \pi^{(1/2)}(a)\big)}{\sigma^2(a)}. \tag{4}$$

When $\Delta(a)$ is positive, the payoff in state 1 is objectively closer to the objective payoff when the agent plays action a; when $\Delta(a)$ is negative, the payoff in state 0 is closer. In particular, when the objective payoff $\tilde{\pi}(a^1)$ is closer to $\pi^1(a^1)$ than to $\pi^0(a^1)$, $\Delta(a_1) > 0$, so that when the agent plays a^1, she becomes more convinced that it is best to play a^1. Similarly, a^0 is attracting if $\tilde{\pi}(a^0)$ is closer to $\pi^0(a^0)$ than to $\pi^1(a^0)$.

Proposition 3. *Fix discount rate r.*

(i) If the interior steady state action exists, it is attracting. In particular, uninformative full-certainty actions are attracting.

(ii) Informative full-certainty action a^1 (a^0) is attracting if $\Delta(a^1) > 0$ ($\Delta(a^0) < 0$) and is repelling if $\Delta(a^1) \le 0$ ($\Delta(a^0) \ge 0$).

(iii) If there are no interior steady state actions and both a^0 and a^1 are repelling, then beliefs and actions converge with probability 0. Otherwise, beliefs and actions converge with probability 1.

The proof is given in the Appendix; here we give the intuition behind the result. Let $[\underline{u}, \overline{u}] \triangleq U$ be the set of interior steady state beliefs. For p slightly outside of U, the volatility of the belief process is separated away from zero and there is a positive chance of hitting U even if the drift leads away from U. Now consider a corner steady state, $p = 1$ for concreteness. The volatility of $\{p_t\}$ vanishes as p_t approaches 1, and we need to look at the drift of the belief process. It turns out that that the sign of the drift is sufficient to determine whether beliefs can converge to 1: positive drift makes p_t converge to 1 with positive probability, while negative drift prevents p_t from converging to 1.

Another intuition involves KL divergence. Consider steady state $p = 1$, and imagine the current belief $p_t \approx 1$. 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's signals on average favor state 1. This makes her willing to keep on playing a^1 and, therefore, she converges to $p = 1$ with positive probability.

The proof works with the agent's belief process and then translates the asymptotic properties of beliefs into the asymptotic properties of actions. We are able to obtain this clean characterization of asymptotic beliefs because diffusion processes in continuous time admit a sharp characterization of the limit distribution (Lemma 5 in the Appendix). However, if the agent only observes the history of the continuous-time process at discrete but sufficiently frequent intervals and changes her actions only at those times, then because the distribution of the agent's future beliefs has no mass points, the process cannot converge to a steady state that is a repellor in the continuous-time model. We believe that attractors of the continuous-time model are also attractors with sufficiently frequent observations in discrete time, but we have not proved this formally.[10] A version of the following result is known in the literature and serves here as an illustration of Propositions 1 and 3: If the support of the agent's prior includes the true state and the agent is sufficiently patient, the limit action is optimal.

Corollary 1. *Assume that a^0 and a^1 are informative, and that $\tilde{\pi} \in \{\pi^0, \pi^1\}$. Then there exists $\bar{r} > 0$ such that for all $r < \bar{r}$, $\lim_{t\to\infty} a_t = \max_a \tilde{\pi}(a)$ with probability 1, so that the agent's action converges to the full-information optimum.*

Proof. Without loss of generality let $\tilde{\pi} = \pi^1$. Because the full-information actions are both informative, Proposition 1 implies that there is a uniform $\bar{r}$ such that for all $p \in [0, 1]$, only informative actions are optimal. Therefore, there are no interior steady states. Because $\tilde{\pi} = \pi^1$, $\Delta(a^1) > 0$, so by Proposition 3, $\tilde{\mathbb{P}}[\lim_{t\to\infty} a_t \to a^1] = 1$. □

The next two corollaries apply for any discount rate $r \in (0, +\infty]$ when the agent's model can be misspecified.

We say that the belief process is *recurrent* on $(0, 1)$ if for every initial belief $p_0 \in (0, 1)$, each belief $p' \in (0, 1)$ is reached with probability 1, i.e.,

$$\tilde{\mathbb{P}}\big[p_t = p' \text{ for some } 0 \le t < \infty\big] = 1.$$

Corollary 2. *Suppose that for all $p \in (0, 1)$, only informative actions are myopically optimal. Then the set of limit beliefs and actions does not depend on r. If in addition a^0 and a^1 are repellors, then the belief interval $(0, 1)$ is recurrent for any r.*

[10]The continuous-time value function is an upper bound on what can be achieved with less frequent observations, and we think it is the limit of the discrete time value functions as the observation periods shrink. Our proof strategy would be to use this to try to show that the optimal actions also converge. Finally, we would need to show that the limiting asymptotic beliefs are the same as the asymptotic beliefs in the continuous-time model.

This follows immediately from Proposition 3 because if all best responses are informative, there are no interior steady states.

We say that there is a *uniform best explanation* if there is a subjective state $\theta \in \{0, 1\}$ such that for all actions $a \in A$,

$$\left|\pi^{\theta}(a) - \tilde{\pi}(a)\right| \leq \frac{1}{2}\left|\pi^{1}(a) - \pi^{0}(a)\right|. \tag{5}$$

The condition posits that the true payoff function $\tilde{\pi}$ is closer to the same subjective payoff π^0 or π^1 for every action. If one state is a strict uniform best explanation (that is, if the inequality in (5) is strict), then the game is weakly identified.

Corollary 3. *Suppose that there is a uniform best explanation. Then the beliefs converge for all r.*

Intuitively, if one subjective state is a better explanation for the agent's observations for every action, then the agent's beliefs get pushed toward that state as long as the agent takes an informative action. Hence, they either converge to probability 1 on that state or get absorbed as soon as the agent takes an uninformative action.

Next, note that if the agent's prior assigns positive probability to the true payoff function $\tilde{\pi}(\cdot)$, then it is the uniform best explanation. Moreover, $\tilde{\pi}(a) + \epsilon(a)$ remains the uniform best explanation if the perturbation $\epsilon(a)$ is sufficiently small for all informative actions and zero for the uninformative action. Hence, in the long run the agent's belief assigns probability 1 to the perturbed version of the the true payoff function.

The next result says that convergence is monotone in the discount rate.

Corollary 4. *If the agent's belief converges for some r', then it converges for any $r'' > r'$.*

Proof. Suppose the belief does not converge for $r'' > r'$. Proposition 3 implies that both a^1 and a^0 are repelling, and there is no interior steady state. Hence, $v_{r''}(p) > g$ for all p. By monotonicity of v_r in r, $v_{r'}(p) \geq v_{r''}(p) > g$ for all p. Therefore, there are no interior steady states under r'. By Proposition 3, the beliefs under r' do not converge, which contradicts the assumption. □

As our leading example in Section 3 shows, the converse is not true: convergence for r' does not imply convergence for $r < r'$.

Corollary 5. *Suppose a^0 and a^1 are repellors, that is, $\Delta(a^1) < 0$ and $\Delta(a^0) > 0$. Then there is $\bar{r}$ such that for all $r < \bar{r}$, the belief interval $(0, 1)$ is recurrent.*

Proof. First, a^0 and a^1 are informative. Indeed, if a^0 and/or a^1 were uninformative, there would be interior steady state beliefs in the neighborhood of the corresponding state, which would be attracting by Proposition 3. Then, by Proposition 1, there is a uniform $\bar{r}$ such that for all $p \in [0, 1]$, only informative actions are optimal, so there are no interior steady states. The only steady state beliefs are $p = 0$ and $p = 1$. The conclusion then follows from Lemma 5 part (i), in the Appendix, which applies a standard result to our setting. □

Corollary 6. *If the uninformative action is myopically optimal for some p, then the beliefs converge for r large enough. If additionally a^0 and a^1 are repellors, then there is an $\bar{r} > 0$ such that for all $r < \bar{r}$, beliefs do not converge.*

Remark 2. Esponda and Pouzo (2016a) (adapted to our one-player setting) say that a game is *weakly identified* at an action a if whenever the two states θ and θ' both minimize the KL divergence between the true and subjective distribution generated by this action, the action generates the same distribution over outcomes in both states. They say that a game is weakly identified if it is weakly identified at every action a. In our model, action a is weakly identified if $\Delta(a) = 0$, which implies that a is uninformative. Their Theorem 4 states that a stable strategy must be a Berk–Nash equilibrium if the game is weakly identified, but inspection of the proof of Theorem 4 shows that for this result it is sufficient to have weak identification of the strategy in question. Similarly, in our model, the game is weakly identified at any attracting steady state: At an interior steady state the action is uninformative and so weakly identified, and if a full certainty action is an attracting steady state, then only the corresponding state minimizes the KL divergence with the truth. A difference between our results is that we characterize the learning dynamics even when the game is not weakly identified at a steady state action a. Proposition 3 part (ii) shows that in this case a is repelling. For instance, if $\tilde{\pi}(a^1) = 0$ and $\pi^1(a^1) = -\pi^0(a^1) > 0$, then a^1 is informative, but $\Delta(a^1) = 0$, so the game is not weakly identified at a^1.

7. Examples

7.1 *Seller with unknown linear demand*

To begin this subsection, we prove Claim 1. Recall

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

The proof is an illustration of how to use the general results of Propositions 1, 2, and 3 in a particular situation. The analysis follows two steps: first we find the set of steady states using Proposition 1 and then we determine whether they are attracting or repelling using Proposition 3. For this example we additionally find the exact discount rate cutoff $\bar{r}$.

Proof of Claim 1. First find that

$$\Delta(a) = -4a\eta/\sigma^2.$$

In this example,

$$a^1 = 1,$$
$$a^0 = -1,$$
$$a^u = 0.$$

We have that $\Delta(a^1) < 0$ and $\Delta(a^0) > 0$. By Proposition 3, both a^1 and a^0 are repelling.

Now we find when the uninformative action $a^u = 0$ is the interior steady state. To do so, we start by computing the value function of the myopic agent:

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

Every optimal myopic strategy satisfies

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

For $\hat{p} \in (1/4, 3/4)$, the only optimal action is uninformative. By Proposition 2, there is $\underline{r}$ such that for $r > \underline{r}$, $a = 0$ is a best response to $\hat{p}$. Thus, $a^u = 0$ is a steady state for $r > \underline{r}$. By Proposition 3, for $r > \underline{r}$, the beliefs and actions converge almost surely.

Conversely, for $\hat{p} > 3/4$ and $\hat{p} < 1/4$, the myopic best response is informative. By Proposition 1, there is $\bar{r}$ such that for all $r < \bar{r}$ and all $p \in [0, 1]$, only informative actions are optimal. Thus in this case, there is no interior steady state. The exact cutoff $\hat{r}$ for the existence/nonexistence of the interior steady state is somewhere in $(\bar{r}, \underline{r})$.

Last, for this example we find the exact value of $\hat{r}$, which here is equal to $6/\sigma^2$. We derive $\hat{r}$ by solving the differential equation for the value function. By Theorem 1, the value function is characterized by the HJB equation

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

If $r < \hat{r}$, we have to have that the maximum in the above expression is attained on $A \setminus \{a^u\} = \{-1, 1\}$. 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

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

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 and C_2 are free constants, and

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

The differentiability of v and symmetry imply 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,

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

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

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

The inequality is satisfied if and only if $r \leq 6/\sigma^2$. □

Now we modify the previous example by supposing that there are only two feasible actions, both of which are informative, so $A = \{-1, 1\}$, as in the example in Section 5 of Nyarko (1991) Note that the uninformative action $a = 0$ is unavailable.

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

Proof. There are no interior steady states because there are no uninformative actions, and both a^1 and a^0 are repelling as shown above. Therefore, the belief and action do not converge by Proposition 3. □

So for any discount rate, the beliefs cycle as in Nyarko's (1991) analysis of the myopic case. Note that both in this two-action example and in our analysis of the three-action case, if $r < \hat{r}$, the action space can effectively be restricted to not contain the uninformative action $a = 0$: in the two-action case, this restriction is exogenous, while in our case with $r < \hat{r}$, the uninformative action is not played because it is never optimal to do so.

7.2 *A bandit model of learning*

Suppose now that there are two actions a^1 and 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 her posterior likelihood is above a threshold $L^\star$. We write $\sigma = \sigma(a^1)$ for the noise level

when the risky arm is chosen. The threshold $L^\star$ depends on s, l, h, and 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 agent's behavior in this example as the agent will stick with the safe arm forever once she has chosen the safe arm for the first time. By Lemma 1, the dynamics of the posterior likelihood of the agent L_t are given by

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

Thus, the posterior likelihood is a Brownian motion with drift $(h-l)(\pi - (h+l)/2)/\sigma^2$ that is absorbed at L^*.

We consider only the nontrivial case when $L_0 > L^*$. From (4) we find that

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

First, a^0 is uninformative and is always a best response for some belief. Therefore, there always is an interior steady state. By Proposition 3, a^0 is attracting. Next, from Proposition 3, a^1 is attracting if $\pi > (h+l)/2$ and repelling if $\pi < (h+l)/2$. 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 strictly positive probability, the agent sticks to the risky arm forever.*

8. 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 depends on the agent's discount rate as well as on the availability of an uninformative action. Our findings have important implications for learning in misspecified games: For play to converge to a Berk–Nash equilibrium, the behavior of the individual players must converge and, as we show, this can depend on the discount rate. Our analysis has been restricted to the case of only two states, as this lets us use existing characterizations of the solutions to one-dimensional SDEs and their limiting behavior (cf. Karatzas and Shreve 1988, Chapter 5). Characterizing the long-run belief dynamics with three or more states would require analogs of those results for higher dimensions.

Appendix A: Appendix: Proofs

A.1 *Proofs omitted from Section 4*

Proof of Lemma 1. Using (1), 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 that determines the true payoff process. As $L(p) = \log(p/(1-p))$ is twice differentiable, we can apply Ito's lemma on (2) and get

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

□

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

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

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 she switches to the optimal uninformative action once her belief reaches an exogenously given set D. The maximal average payoff from playing the uninformative action forever equals

$$g \triangleq \pi^0(a^u).$$

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 $a_t \in A \setminus a^u$, admits a value function that is twice differentiable on $[0,1] \setminus D$ and solves

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

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

Proof. We use Strulovici and Szydlowski (2015) to establish the result. As the variance of the belief process is not uniformly bounded from below, the conditions of Strulovici and Szydlowski (2015) are not satisfied directly.[11] To avoid this problem, we use the log-likelihood ratio process L_t instead of the belief process p_t as a state variable:

$$L_t \triangleq \log \frac{p_t}{1-p_t}.$$

We denote by $w(L) \triangleq v_r(p(L))$ the value function of the agent when she holds the belief

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

For further reference, let us note that

$$p'(L) = \frac{e^L}{\left(1+e^L\right)^2} = \left(1-p(L)\right)p(L), \tag{A.1}$$

$$p''(L) = p'(L)\left(1-2p(L)\right). \tag{A.2}$$

The dynamics of L_t are given by

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

where W_t is a Brownian motion according to the agent's subjective probability measure. Note that the drift term $\mu(L,a) \triangleq [p(L_t) - 1/2]I(a)_t^2$ is bounded by

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

the volatility is bounded from above and below by

$$\min_{a' \in A\setminus a^u} I\left(a'\right) \leq I(a) \leq \max_{a' \in A\setminus a^u} I\left(a'\right),$$

and the flow payoffs are bounded by

$$\left|\pi_{p(L)}(a)\right| \leq \max_a \max\left\{\left|\pi^1(a)\right|, \left|\pi^0(a)\right|\right\}.$$

Furthermore, the variance term is independent of L and thus is Lipschitz continuous, the flow payoff and the drift of L are linear in the belief p, and as the belief is Lipschitz continuous in L it is Lipschitz continuous in L as well. Thus, Assumptions 1, 2, and 3 from Strulovici and Szydlowski (2015) are satisfied, and it follows that the value function $w : \mathbb{R} \to \mathbb{R}$ is twice differentiable and satisfies

$$w(L) = \max_{a \in A\setminus a^u} \pi_{p(L)}(a) + \frac{I(a)^2}{r}\left(\left[p(L) - \frac{1}{2}\right]w'(L) + \frac{1}{2}w''(L)\right). \tag{A.3}$$

[11] 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.

Note that as $w(L) = v_r(p(L))$ and by (A.1) and (A.2), we have that

$$\begin{aligned} w'(L) &= v_r'\big(p(L)\big)p'(L) = v_r'(p)p(1-p), \\ w''(L) &= v_r''\big(p(L)\big)\big[p'(L)\big]^2 - p''(L)v_r'(L) \\ &= v_r''(p)\big[p(1-p)\big]^2 - 2\left[p - \frac{1}{2}\right]w'(L). \end{aligned}$$

Plugging this into (A.3) yields

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

THEOREM 2. *For each discount rate $r \in (0, +\infty]$, there exists an optimal strategy $s_r^* : [0,1] \to A$, which is Markovian. The value function $v_r : [0,1] \to \mathbb{R}_+$ is continuous, convex in the belief p, and continuous in the discount rate r. Furthermore, $v_r(p)$ is nonincreasing in r and $\lim_{r\to\infty} v_r(p) = v_\infty(p)$ for all p. Furthermore, for $r < \infty$, it is twice continuously differentiable on $\{p \in (0,1) : v_r(p) > g\}$ and satisfies the HJB equation*

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

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

PROOF. First note that v_r is well defined, as an upper bound on v_r is the payoff from taking the optimal action forever,

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

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

$$v_r(p) \ge \max_a \pi^{(p)}(a).$$

As v_r is the supremum over linear functions, it is convex. As every convex function is continuous, v_r is continuous and the set of beliefs $U \subseteq [0,1]$ for which an uninformative action is optimal is closed:

$$U \triangleq \big\{p \in [0,1] : g = v_r(p)\big\}.$$

As U is closed, we can define the first time the belief process reaches U, $\tau_U \triangleq \inf\{t : p_t \in U\}$. Define $\hat{x}$ as the process that is absorbed in U:

$$\hat{x}_t = x_{\min\{t, \tau_U\}}.$$

Consider the control problem where the agent is restricted to informative controls $a_t \in A \setminus a^u$ and the process is absorbed the first time it leaves $[0,1] \setminus U$ with a payoff of g.

By Lemma 3, the value function $\hat{v}_r : [0,1] \setminus U \to \mathbb{R}$ of this problem is twice differentiable and solves

$$\hat{v}_r(p) = \max_{a \in A \setminus a^u} \pi^{(p)}(a) + \big[I(a)p(1-p)\big]^2 \frac{\hat{v}_r''(p)}{2r}, \tag{A.4}$$

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

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

$$\begin{aligned} v_r(p) &= r\mathbb{E}\bigg[\int_0^{\hat{\tau}} e^{-rt}\pi^{(p_t)}(a_t)\,\mathrm{d}t + \int_{\hat{\tau}}^{\infty} e^{-rt}\pi^{(p_t)}(a_t)\,\mathrm{d}t \;\bigg|\; p_0 = p\bigg] \\ &= \big(1 - \mathbb{E}\big[e^{-r\hat{\tau}}\big]\big)g(p) + \mathbb{E}\big[e^{-r\hat{\tau}}\big]v_r(p) \\ \Rightarrow \quad 0 &= \big(1 - \mathbb{E}\big[e^{-r\hat{\tau}}\big]\big)\big(g(p) - v_r(p)\big). \end{aligned}$$

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

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

$$\begin{aligned} &r\mathbb{E}\bigg[\int_0^{\infty} e^{-rt}\pi^{(p)}(a_t)\,\mathrm{d}t\bigg] \\ &\quad = r\mathbb{E}\bigg[\int_0^{\tau_U} e^{-rt}\pi^{(p_t)}(a_t)\,\mathrm{d}t + \mathbb{E}\bigg[\int_{\tau_U}^{\infty} e^{-rt}\pi^{(p_t)}(a_t)\,\mathrm{d}t \;\bigg|\; p_{\tau_U}\bigg]\bigg] \\ &\quad \le r\mathbb{E}\bigg[\int_0^{\tau_U} e^{-rt}\pi^{(p_t)}(a_t)\,\mathrm{d}t + e^{-r\tau_U} g\bigg]. \end{aligned} \tag{A.5}$$

In the next step, we use that by (A.4), for all $p \in [0,1] \setminus U$ and all $a \neq a^u$, we have

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

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

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

By Ito's lemma and Doob's optional sampling theorem, we have that

$$\begin{aligned}
&\mathbb{E}\left[\int_0^{\tau_U} re^{-rt}\pi^{(p_t)}(a_t)\,\mathrm{d}t\right] \\
&\quad\le \mathbb{E}\left[\int_0^{\tau_U}\left\{re^{-rt}\hat{v}_r(p) - e^{-rt}\big[I(a)p(1-p)\big]^2\frac{\hat{v}_r''(p)}{2}\right\}\mathrm{d}t\right] \\
&\quad= \mathbb{E}\big[\hat{v}_r(p_0) - e^{-rp_{\tau_U}}\hat{v}_r(p_{\tau_U})\big].
\end{aligned} \tag{A.6}$$

Combining (A.5) and (A.6) and the fact that $\hat{v}_r(p_{\tau_U}) = g$ by the boundary condition of (A.4) yields

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

Thus, for any policy a, the value is lower than the value when following the policy $a^\star$. This shows that the HJB equation describes an optimal policy. Moreover, it is clear from inspecting the HJB equation that multiplying the discount rate by $\lambda > 1$ is the same as multiplying the volatility $\sigma(\cdot)$ by $\sqrt{\lambda}$, which is the same as adding noise to the agent's signals. Hence, the agent's value function is decreasing in r.

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

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

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_r 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

$$v_r(p) \le p\pi^0\big(a^0\big) + (1-p)\pi^1\big(a^1\big).$$

The agent can take an informative action for a long, but deterministic, time T to learn the state arbitrarily precisely and afterward take an optimal action. As the agent becomes patient, her 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

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

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

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

To see that $v_\infty(p)$ is also an upper bound, observe that the agent's continuation payoff after time τ cannot be better than when learning the state at time τ for free and taking the subjectively optimal action at all future points in time:

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

As $\pi^{(p_t)}(a_t) \leq \max_a \pi^{(p_t)}(a) = v_\infty(p_t)$, we have that for every fixed strategy s,

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

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

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

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

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

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

$$\lim_{r\to\infty} \mathbb{E}^s\Big[\sup_{t\in[0,1/\sqrt{r}]} v_\infty(p_t) \;\Big|\; p_0 = p\Big] = \mathbb{E}^s\Big[\lim_{r\to\infty} \sup_{t\in[0,1/\sqrt{r}]} v_\infty(p_t) \;\Big|\; p_0 = p\Big] = v_\infty(p),$$

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

Lemma 4. *Fix an arbitrary strategy s and any $\underline{p} \in (0,1)$, and let $\tau_{(\underline{p},1)}$ be the first time that the belief leaves the interval $(\underline{p},1)$. Then*

$$\lim_{p_0 \nearrow 1} \mathbb{E}^s\big[e^{-r\tau_{(\underline{p},1)}} \mid \theta = 1\big] = 0.$$

Proof. Note that $e^{-r\tau_{(\underline{p},1)}} = 0$ whenever $\lim_{t\to\infty} p_{t\wedge\tau_{(\underline{p},1)}} = 1$. We show that this event happens with probability 1 as $p_0 \to 1$. Because the belief conditional on $\theta = 1$ is a submartingale under any strategy and τ is a stopping time, the fact that $p_{\tau_{(\underline{p},1)}}$ is equal either to 1 or to $\underline{p}$ implies that

$$p_0 \le \mathbb{E}^s[p_{\tau_{(\underline{p},1)}} \mid \theta = 1] = \mathbb{P}^s[p_{\tau_{(\underline{p},1)}} = 1 \mid \theta = 1] + \underline{p}\big\{1 - \mathbb{P}^s[p_{\tau_{(\underline{p},1)}} = 1 \mid \theta = 1]\big\},$$

as

$$1 \le \Big(\lim_{p_0\to 1} \mathbb{P}^s[p_{\tau_{(\underline{p},1)}} = 1 \mid \theta = 1]\Big)(1 - \underline{p}) + \underline{p}$$

$$\Leftrightarrow \quad 1 \le \lim_{p_0\to 1} \mathbb{P}^s[p_{\tau_{(\underline{p},1)}} = 1 \mid p_0 = \tilde{p}, \theta = 1].$$

Consequently, we have that $\lim_{p_0\to 1} \mathbb{P}^s[p_{\tau_{(\underline{p},1)}} = 1 \mid \theta = 1] = 1$ and, hence,

$$\lim_{p_0\nearrow 1} \mathbb{E}^s\big[e^{-r\tau_{(\underline{p},1)}} \mid \theta = 1\big] \le \Big(1 - \lim_{p_0\nearrow 1} \mathbb{P}^s[p_{\tau_{(\underline{p},1)}} = 1 \mid \theta = 1]\Big) = 0. \qquad \square$$

A.2 *Proofs omitted from Section 5*

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, $v_r(p') \ge g$. But convexity of v_r implies that $v_r(p') \le g$ 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 $a, a' \in A \setminus a^u$ 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 ordinary differential equations (ODEs) on $[p', p'']$:

$$v_r(p) = \pi^{(p)}(a) + \big[I(a)p(1-p)\big]^2 \frac{v_r''(p)}{2r}, \tag{A.7}$$

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

Take the difference between these ODEs to find

$$0 = \pi^{(p)}(a) - \pi^{(p)}(a') + \big[p(1-p)\big]^2 \frac{v_r''(p)}{2r}\big(I(a)^2 - I(a')^2\big). \tag{A.8}$$

As $\pi^{(p)}(\cdot)$ is linear in p and as there are no informationally equivalent actions, $I(a) \neq I(a')$, it follows from (A.8) that $[p(1-p)]^2 v_r''(p)/(2r)$ must be linear in p. By (A.7), $v_r(p)$

is hence linear in p. This implies that $v_r''(p) = 0$ and by (A.8), we have that for all $p \in [p', p'']$,

$$\pi^{(p)}(a) = \pi^{(p)}(a') \quad \Rightarrow \quad a = a'.$$

Finally, as the informative action is almost everywhere unique and the solution to the SDE of the belief process (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.

We finally show the third part of the lemma. First, we assume that the action a^1 is optimal for some belief in every neighborhood of $p = 1$, that is, that for all $p' < 1$, there is $p'' \in (p', 1)$ such that a^1 is optimal at p''. In this case, we say that a^1 *has nontrivial support*. We show that no other action $\hat{a}$ can be optimal in every neighborhood of $p = 1$. Consider two subcases: a^1 is either informative or uninformative.

Case 1: a^1 has nontrivial support and $I(a^1) \neq 0$. Suppose, to the contrary, that in any neighborhood of $p = 1$, there is a belief such that some action different from a^1 is optimal. Because v_r is continuous in p, this implies that there must be a sequence of beliefs $(\tilde{p}^j)_j \to 1$ such that at each $\tilde{p}^j$, the agent is indifferent between a^1 and some other action. As there are only finitely many actions, there must then be an action $\hat{a}$ and subsequence $(j_k)_k$ with $\lim_{k\to\infty} j_k = \infty$ such that the agent is indifferent between a^1 and $\hat{a}$ at all beliefs $(\tilde{p}^{j_k})_{k\in\{1,2,\dots\}}$. Denote $p^k = \tilde{p}^{j_k}$.

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

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

$$v_r(p^k) = \pi^{p^k}(\hat{a}) + \big(I(\hat{a})p^k(1-p^k)\big)^2 \frac{v_r''(p^k)}{2r}.$$

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

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

However, the second equality implies that

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

Taking the ratio, we obtain

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

which is a contradiction to the assumption that $I(a^1) \neq 0$.

Case 2: a^1 has nontrivial support and $I(a^1) = 0$. By the first part of the lemma, we know that a^1 is optimal on some interval $[\underline{u}, \overline{u}] \subset [0, 1]$. As a^1 is optimal for some belief

in every neighborhood of $p = 1$, it follows that the action a^1 is optimal for some interval $[\underline{u}, 1]$. But then, by the proof for Case 1, no other action is optimal in $[\underline{u}, 1]$.

Case 3. Finally, we deal with the case that the action a^1 is optimal for $p = 1$, but has trivial support, so there is a $\underline{p} < 1$such that a^1 is not optimal for any $p \in (\underline{p}, 1)$. The agent's value when she uses the strategy s is bounded above by the payoff she gets when she takes the correct action a^0 in the low state $\theta = 0$ and uses the strategy s in the high state:

$$\mathbb{E}^s\left[\int_0^\infty e^{-rt}\,\mathrm{d}\pi_t\right] \le (1-p)\pi^0(a^0) + p\mathbb{E}^s\left[\int_0^\infty e^{-rt}\,\mathrm{d}\pi_t \,\middle|\, \theta = 1\right].$$

Let $\tau_{(\underline{p},1)}$ be the first time that the belief reaches either $\underline{p}$ or 1. Then since the agent does not play a_1 until at least time $\tau_{(\underline{p},1)}$,

$$\begin{aligned}\mathbb{E}^s\left[\int_0^\infty e^{-rt}\,\mathrm{d}\pi_t \,\middle|\, \theta = 1\right] \le{}& \left(1 - \mathbb{E}^s\left[e^{-r\tau_{(\underline{p},1)}} \mid \theta = 1\right]\right)\left(\max_{a\neq a^1} \pi^1(a)\right)\\ &+ \mathbb{E}^s\left[e^{-r\tau_{(\underline{p},1)}} \mid \theta = 1\right]\pi^1(a^1).\end{aligned}$$

Combining this with the previous inequality gives

$$\begin{aligned}\mathbb{E}^s\left[\int_0^\infty e^{-rt}\,\mathrm{d}\pi_t\right] \le{}& (1-p)\pi^0(a^0)\\ &+ p\left\{\left(1 - \mathbb{E}^s\left[e^{-r\tau_{(\underline{p},1)}} \mid \theta = 1\right]\right)\left(\max_{a\neq a^1} \pi^1(a)\right)\right.\\ &\left.+ \mathbb{E}^s\left[e^{-r\tau_{(\underline{p},1)}} \mid \theta = 1\right]\pi^1(a^1)\right\}.\end{aligned} \tag{A.9}$$

As shown in Lemma 4, the expected discounted time until the belief leaves the interval $(\underline{p}, 1)$ conditional on the high state $\theta = 1$ goes to zero when the initial belief p_0 goes to 1 for every strategy, i.e.,

$$\lim_{p_0 \nearrow 1} \mathbb{E}^s\left[e^{-r\tau_{(\underline{p},1)}} \mid \theta = 1\right] = 0.$$

Consequently, taking the limit $p_0 \nearrow 1$ of (A.9) yields that

$$\lim_{p_0\nearrow 1} v_r(p_0) = \lim_{p_0\nearrow 1} \mathbb{E}^s\left[\int_0^\infty e^{-rt}\,\mathrm{d}\pi_t\right] \le \max_{a\neq a^1} \pi^1(a) < \pi^1(a^1) = v_r(1).$$

This shows that for a^1 not to be optimal in the interval $(\underline{p}, 1)$, the agent's payoff when she is almost certain that the state is 1 must be bounded away from the payoff she gets when she knows the state is 1. This contradicts the continuity of v_r.

Hence, we have shown that the action a^1 is the unique optimal action for a nonempty interval of beliefs $[\underline{p}, 1)$. By the analogous argument, it follows that the action a^0 is the unique optimal action for a nonempty interval of beliefs around $p = 0$. □

Proposition 1. *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 $r < \bar{r}$ and all $p \in [0, 1]$, only informative actions are optimal.*

Proof. The case where all actions are informative is trivial. Suppose there is an uninformative action a^u and recall that g is the payoff from playing a^u forever. We first show that our assumptions on payoff functions imply that there exists a belief $\hat{p} \in (0,1)$ such that all myopic best responses to $\hat{p}$ are informative. Suppose, to the contrary, that the uninformative action a^u is myopically optimal for all $p \in (0,1)$. Then Lemma 2 implies that a^u is optimal for $p \in \{0,1\}$, which contradicts the assumption that the full-certainty actions are different.

We just showed 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 thereafter. The value function from following this strategy is denoted by $\tilde{V}_r(p)$. For any $p \in (0,1)$, we have

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

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 $p_0 \in (0,1)$, the distribution of p_τ has full support on $[0,1]$ and so $\mathbb{E}[v_\infty(p_\tau) - g] > 0$. As $\mathbb{E}[v_\infty(p_\tau) - g \mid p_0 = p]$ 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 $p \mapsto \mathbb{E}[v_\infty(p_\tau) - g \mid p_0 = p]$ is bounded away from zero, as it is continuous and strictly positive. Therefore, there is uniform $\bar{r}$ such that $\tilde{V}_r(p) > g$ 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. □

A.3 *Proofs omitted from Section 6*

Let the diffusion process L be defined on $(\underline{L}, \overline{L})$, $-\infty \leq \underline{L} < \overline{L} \leq +\infty$, by

$$dL_t = \alpha(L_t)\,dt + \beta(L_t)\,dW_t, \tag{A.10}$$

where Borel measurable coefficients $\alpha, \beta : \mathbb{R} \to \mathbb{R}$ satisfy

$$\beta(L) \neq 0, \quad \forall x \in (\underline{L}, \overline{L})$$

$$\forall x \in (\underline{L}, \overline{L}), \exists \varepsilon > 0 \quad \text{such that} \quad \int_{x-\varepsilon}^{x+\varepsilon} \frac{1 + \big|\alpha(y)\big|}{\beta(y)^2}\,dy < \infty.$$

Fix an arbitrary $L_0 \in (\underline{L}, \overline{L})$. The *scale function* for L is a (strictly increasing and invertible) function $\phi : (\underline{L}, \overline{L}) \to \mathbb{R}$ defined by

$$\phi(L) = \int_{L_0}^{L} \exp\left(-\int_{L_0}^{y} \frac{2\alpha(z)}{\beta(z)^2}\,\mathrm{d}z\right)\mathrm{d}y. \tag{A.11}$$

The next lemma is Proposition 5.22 in Karatzas and Shreve (1988, p. 345) written in our notation.

Lemma 5. *Let $T = \inf\{t \geq 0 : L_t \notin (\underline{L}, \overline{L})\}$ be the exit time from $(\underline{L}, \overline{L})$. Every weak solution of (A.10) has the following properties:*

(i) If $\phi(\underline{L}+) = -\infty$ and $\phi(\overline{L}-) = \infty$, then

$$\mathbb{P}[T = \infty] = \mathbb{P}\Big[\sup_{0 \leq t < \infty} L_t = \overline{L}\Big] = \mathbb{P}\Big[\inf_{0 \leq t < \infty} L_t = \underline{L}\Big] = 1.$$

In particular, the process L is recurrent: for every $y \in (\underline{L}, \overline{L})$, we have

$$\mathbb{P}[L_t = y \text{ for some } 0 \leq t < \infty] = 1.$$

(ii) If $\phi(\underline{L}+) > -\infty$ and $\phi(\overline{L}-) = \infty$, then the process is absorbed in $\underline{L}$ with probability 1:

$$\mathbb{P}\Big[\lim_{t \to T} L_t = \underline{L}\Big] = 1.$$

(iii) If $\phi(\underline{L}+) = -\infty$ and $\phi(\overline{L}-) < \infty$, then the process is absorbed in $\overline{L}$ with probability 1:

$$\mathbb{P}\Big[\lim_{t \to T} L_t = \overline{L}\Big] = 1.$$

(iv) If $\phi(\underline{L}+) > -\infty$ and $\phi(\overline{L}-) < \infty$, then the probability that the process is absorbed in $\overline{L}$ ($\underline{L}$) is given by

$$\mathbb{P}\Big[\lim_{t \to T} L_t = \underline{L}\Big] = 1 - \mathbb{P}\Big[\lim_{t \to T} L_t = \overline{L}\Big] = \frac{\phi(\overline{L}-) - \phi(L_0)}{\phi(\overline{L}-) - \phi(\underline{L}+)}.$$

Proposition 3. *Fix discount rate r.*

(i) If the interior steady state action exists, it is attracting. In particular, uninformative full-certainty actions are attracting.

(ii) Informative full-certainty action a^1 (a^0) is attracting if $\Delta(a^1) > 0$ ($\Delta(a^0) < 0$) and is repelling if $\Delta(a^1) \leq 0$ ($\Delta(a^0) \geq 0$).

(iii) If there are no interior steady state actions and both a^0 and a^1 are repelling, then beliefs and actions converge with probability 0. Otherwise, beliefs and actions converge with probability 1.

Proof. Let $\underline{p}$ be the largest steady state belief below p_0 and let $\overline{p}$ be the smallest steady state belief above p_0. If p_0 is already a steady state, then the analysis is trivial. In what follows, we study the nontrivial case. Change the state variable from p to $L = \log(p/(1-p))$ and find $\overline{L}$ and $\underline{L}$ that 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 that says that to get the limit distribution of the process, it is sufficient to evaluate the natural scale function (A.11) at boundaries $\underline{L}$ and $\overline{L}$. Fix arbitrary z and consider

$$\phi_r(\overline{L}) = \int_{L_0}^{\overline{L}} \exp\left\{-\int_z^x \frac{2\alpha(y)}{\beta^2(y)}\,dy\right\} dx,$$
$$\phi_r(\underline{L}) = \int_{L_0}^{\underline{L}} \exp\left\{-\int_z^x \frac{2\alpha(y)}{\beta^2(y)}\,dy\right\} dx,$$

where $\alpha(L) = I(a)(\tilde{\pi}(a) - \pi^{(1/2)}(a))/\sigma(a) = \Delta(a)$ is the drift of L_t, where $a = s_r^*(L)$, and $\beta(L) = I(a)$ is the volatility of L_t; see (3). Therefore,

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

By definition of $\underline{L}$ and $\overline{L}$, $\beta(L)^2 > 0$ for all $L \in (\underline{L}, \overline{L})$. There are four cases to consider: $\overline{L} = \infty$, $\underline{L} = -\infty$, $\overline{L} < \infty$, and $\underline{L} > -\infty$.

We start with $\overline{L} = +\infty$. By Lemma 2, there is $p^* < 1$ such that a^1 is optimal for all $p \in [p^*, 1]$. Write L^* for the corresponding log-likelihood ratio. We have

$$\begin{aligned}
\phi_r(+\infty) &= \int_{L_0}^{+\infty} \exp\left\{-\int_z^x \frac{2\alpha(y)}{\beta^2(y)}\,dy\right\} dx \\
&= K_1 + \int_{L^*}^{+\infty} \exp\left\{-\int_z^x \frac{2\alpha(y)}{\beta^2(y)}\,dy\right\} dx \\
&= K_1 + \int_{L^*}^{+\infty} \exp\left\{-\int_z^x \frac{2\Delta(a^1)}{I(a^1)^2}\,dy\right\} dx \\
&= K_1 + \int_{L^*}^{+\infty} \exp\left\{-(x-z)\frac{2\Delta(a^1)}{I(a^1)^2}\right\} dx \\
&= K_1 + K_2 \int_{L^*}^{+\infty} \exp\left\{-x\frac{2\Delta(a^1)}{I(a^1)^2}\right\} dx
\end{aligned} \tag{A.12}$$

for some finite K_1 and $0 < K_2 < \infty$. Now if $\Delta(a^1) > 0$, then the integral in (A.12) converges, and $\phi_r(+\infty) < \infty$. Conversely, if $\Delta(a^1) \le 0$, then it diverges and $\phi_r(+\infty) = \infty$.

Therefore, we have

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

By Lemma 5, $L = +\infty$ is attracting if $\phi_r(+\infty) < +\infty$ and repelling if $\phi_r(+\infty) = +\infty$. The case $\underline{L} = \infty$ is done in a similar fashion. We have

$$\phi_r(-\infty) \begin{cases} > -\infty & \text{if } \Delta(a^0) < 0, \\ = -\infty & \text{if } \Delta(a^0) \geq 0. \end{cases}$$

By Lemma 5, $L = -\infty$ is attracting if $\phi_r(-\infty) > -\infty$ and repelling if $\phi_r(-\infty) = -\infty$.

Now let us do the case $\overline{L} < \infty$. Since $\Delta(a)$ is bounded and $I(s_r^*(L))$ is bounded away from zero on $L \in (\underline{L}, \overline{L})$, $2\alpha(L)/\beta^2(L)$ is bounded on the bounded interval $[L_0, \overline{L}]$. Therefore, $\phi_r(\overline{L}) < \infty$. By Lemma 5, $\overline{L}$ is attracting.

Absolutely analogously find that $\phi_r(\underline{L}) < \infty$ and $\underline{L}$ is attracting.

Finally, Lemma 5 implies that convergence has a 0–1 property. Therefore part (iii) of the proposition follows. □

Proof of Corollary 3. Suppose condition (5) holds for $\theta = 1$, $|\pi^1(a) - \tilde{\pi}(a)| \leq (1/2)|\pi^1(a) - \pi^0(a)|$, $\forall a \in A$. Then

$$\pi^1(a) - \frac{1}{2}\left|\pi^1(a) - \pi^0(a)\right| \leq \tilde{\pi}(a) \leq \pi^1(a) + \frac{1}{2}\left|\pi^1(a) - \pi^0(a)\right|,$$

$$\begin{aligned} \pi^1(a) - \pi^0(a) - \frac{1}{2}\left|\pi^1(a) - \pi^0(a)\right| &\leq \tilde{\pi}(a) - \pi^0(a) \\ &\leq \pi^1(a) - \pi^0(a) + \frac{1}{2}\left|\pi^1(a) - \pi^0(a)\right|. \end{aligned}$$

If $\pi^1(a) - \pi^0(a) > 0$, then

$$\frac{1}{2}\left|\pi^1(a) - \pi^0(a)\right| \leq \tilde{\pi}(a) - \pi^0(a) \leq \frac{3}{2}\left|\pi^1(a) - \pi^0(a)\right|.$$

If $\pi^1(a) - \pi^0(a) < 0$, then

$$-\frac{3}{2}\left|\pi^1(a) - \pi^0(a)\right| \leq \tilde{\pi}(a) - \pi^0(a) \leq -\frac{1}{2}\left|\pi^1(a) - \pi^0(a)\right|.$$

Putting the two cases together, we find that

$$\left|\pi^0(a) - \tilde{\pi}(a)\right| \geq \frac{1}{2}\left|\pi^1(a) - \pi^0(a)\right| \quad \forall a \in A.$$

Now

$$\Delta(a) = \frac{\left(\pi^0(a) - \tilde{\pi}(a)\right)^2 - \left(\pi^1(a) - \tilde{\pi}(a)\right)^2}{\sigma^2(a)} \geq 0 \quad \forall a.$$

By the second part of Proposition 3, a^1 is attracting. By the third part of Proposition 3, the belief converges. □

References

Aghion, Philippe, Patrick Bolton, Christopher Harris, and Bruno Jullien (1991), "Optimal learning by experimentation." *The Review of Economic Studies*, 58, 621–654. [1156]

Arrow, Kenneth and Jerry Green (1973), "Notes on expectations equilibria in Bayesian settings." Working Paper 33, Institute for Mathematical Studies in the Social Sciences, Stanford University. [1157]

Arrow, Kenneth J., David Blackwell, and Meyer A. Girshick (1949), "Bayes and minimax solutions of sequential decision problems." *Econometrica*, 17, 213–244. [1156]

Berk, Robert (1966), "Limiting behavior of posterior distributions when the model is incorrect." *The Annals of Mathematical Statistics*, 37, 51–58. [1157, 1164]

Bolton, Patrick and Christopher Harris (1999), "Strategic experimentation." *Econometrica*, 67, 349–374. [1162]

Chernoff, Herman (1972), *Sequential Analysis and Optimal Design*, volume 8 of CBMS–NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, Pennsylvania. [1156]

Cho, In-Koo and Kenneth Kasa (2015), "Learning and model validation." *The Review of Economic Studies*, 82, 45–82. [1162]

Diaconis, Persi and David Freedman (1986), "On the consistency of Bayes estimates." *The Annals of Statistics*, 14, 1–26. [1156]

Dubins, Lester E. and David A. Freedman (1966), "Invariant probabilities for certain Markov processes." *The Annals of Mathematical Statistics*, 37, 837–848. [1157]

Easley, David and Nicholas M. Kiefer (1988), "Controlling a stochastic process with unknown parameters." *Econometrica*, 56, 1045–1064. [1156]

Esponda, Ignacio (2008), "Behavioral equilibrium in economies with adverse selection." *American Economic Review*, 98, 1269–1291. [1157]

Esponda, Ignacio and Demian Pouzo (2016a), "Berk–Nash equilibrium: A framework for modeling agents with misspecified models." *Econometrica*, 84, 1093–1130. [1157, 1158, 1162, 1170]

Esponda, Ignacio and Demian Pouzo (2016b), "Equilibrium in misspecified Markov decision processes." Unpublished paper, University of California, Berkeley. [1158]

Eyster, Erik and Matthew Rabin (2005), "Cursed equilibrium." *Econometrica*, 73, 1623–1672. [1157]

Fudenberg, Drew and David M. Kreps (1993), "Learning mixed equilibria." *Games and Economic Behavior*, 5, 320–367. [1158, 1162]

Fudenberg, Drew and David M. Kreps (1995), "Learning in extensive-form games I. Self-confirming equilibria." *Games and Economic Behavior*, 8, 20–55. [1156]

Fudenberg, Drew and David K. Levine (1993a), "Self-confirming equilibrium." *Econometrica*, 61, 523–545. [1156]

Fudenberg, Drew and David K. Levine (1993b), "Steady state learning and Nash equilibrium." *Econometrica*, 61, 547–573. [1156]

Fudenberg, Drew, Philipp Strack, and Tomasz Strzalecki (2017), "Stochastic choice and optimal sequential sampling." Unpublished paper, Harvard University. [1156]

Gallant, A. Ronald and Halbert White (1988), *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Blackwell, New York. [1157]

Gittins, John C. (1979), "Bandit processes and dynamic allocation indices." *Journal of the Royal Statistical Society, Series B, Methodological*, 41, 148–177. [1156]

Heidhues, Paul, Botond Kőszegi, and Philipp Strack (2017), "Unrealistic expectations and misguided learning." Unpublished paper, Central European University. [1158]

Jehiel, Philippe (2005), "Analogy-based expectation equilibrium." *Journal of Economic Theory*, 123, 81–104. [1157]

Jehiel, Philippe and Frédéric Koessler (2008), "Revisiting games of incomplete information with analogy-based expectations." *Games and Economic Behavior*, 62, 533–557. [1157]

Karatzas, Ioannis and Steven E. Shreve (1988), *Brownian Motion and Stochastic Calculus*, volume 113 of Graduate Texts in Mathematics. Springer-Verlag, New York. [1173, 1184]

Kiefer, Nicholas M. and Yaw Nyarko (1989), "Optimal control of an unknown linear process with learning." *International Economic Review*, 30, 571–586. [1156]

Liptser, Robert and Albert N. Shiryaev (1974), *Statistics of Random Processes: I. General Theory*. Nauka, Moscow. [1162]

McLennan, Andrew (1984), "Price dispersion and incomplete learning in the long run." *Journal of Economic Dynamics and Control*, 7, 331–347. [1156]

Moscarini, Giuseppe and Lones Smith (2001), "The optimal level of experimentation." *Econometrica*, 69, 1629–1644. [1156]

Nyarko, Yaw (1991), "Learning in mis-specified models and the possibility of cycles." *Journal of Economic Theory*, 55, 416–427. [1157, 1160, 1172]

Sargent, Thomas J. (1999), *The Conquest of American Inflation*. Princeton University Press, Princeton, New Jersey. [1162]

Shalizi, Cosma Rohilla (2009), "Dynamics of Bayesian updating with dependent data and misspecified models." *Electronic Journal of Statistics*, 3, 1039–1074. [1157]

Strulovici, Bruno H. and Martin Szydlowski (2015), "On the smoothness of value functions and the existence of optimal strategies." *Journal of Economic Theory*, 159, 1016–1055. [1164, 1175]

Wald, Abraham (1945), "Sequential tests of statistical hypotheses." *The Annals of Mathematical Statistics*, 16, 117–186. [1156]

Whittle, Peter (1980), "Multi-armed bandits and the Gittins index." *Journal of the Royal Statistical Society, Series B, Methodological*, 42, 143–149. [1156]

Co-editor Johannes Hörner handled this manuscript.

Manuscript received 24 March, 2016; final version accepted 13 November, 2016; available online 5 December, 2016.