Limits Points of Endogenous Misspecified Learning

Drew Fudenberg†  Giacomo Lanzani‡  Philipp Strack §

First posted version: March 10, 2020
This version: March 14, 2020

Abstract

We study how a misspecified agent learns from endogenous data when their prior belief does not impose restrictions on the distribution of outcomes, but can assign probability 0 to a neighborhood of the true model. We characterize which actions are stable, in the sense that they have a very high probability of being the long-run outcome for some initial beliefs, and which are positively attracting, meaning that they have positive probability of being the long-run outcome for any initial full support belief. Our characterizations are based on two variants of Berk-Nash equilibria: A Berk-Nash equilibrium is uniformly strict if the equilibrium action is a strict best response to all the outcome distributions that minimize the Kullback-Leibler divergence from the truth, and uniform if the action is a best response to all those distributions. Uniformly strict Berk-Nash equilibria are stable, and only uniform Berk-Nash equilibria are stable. All the uniformly strict Berk-Nash equilibria are positively attracting under causation neglect, where the agent believes that their action does not influence the outcome, and under correlation neglect, where the agent believes that the outcome distribution associated with one action does not convey information about the outcomes associated with other actions. In supermodular decision problems, extremal actions are positively attracting if they are uniformly strict Berk-Nash equilibria. We generalize some results to settings where the agent observes a signal before choosing their action.

*We thank Demian Pouzo, Ignacio Esponda, Kevin He, Mira Frick, Roberto Corrao, Ryota Iijima, Yuhta Ishi, and Yuichi Yamamato for helpful conversations, and National Science Foundation grant SES 1643517 and the Guido Cazzavillan Scholarship for financial support.
†Department of Economics, MIT
‡Department of Economics, MIT
§Department of Economics, Yale University
1 Introduction

We study the joint evolution of an agent’s actions and beliefs when their action can influence the distribution of outcomes, and their prior may be misspecified in the sense that it assigns probability 0 to a neighborhood of the true model. We assume that the sets of actions and outcomes are finite, but do not restrict the agent’s prior beliefs to a finite number of states or any specific functional form. In this general environment, the agent’s prior is a belief over the set of action-contingent outcome distributions, and the agent is misspecified if they assign probability 0 to a neighborhood of the true distribution of outcomes for some action. The agent’s prior also determines how they perceive the correlation between the outcome distributions induced by different actions, which we show is a key determinant of the long-run outcome of the learning process.

Our results characterize the possible limit points of the agent’s action and their stability properties. First, Proposition 1 shows that if the agent plays myopically and play converges to an action $a$, then it converges to a Berk-Nash equilibrium (Esponda and Pouzo, 2016), and in addition that if the agent’s prior has finite support then the limit action must be a uniform Berk-Nash equilibrium. Berk-Nash equilibrium requires that the action is myopically optimal against some belief that is supported over the action-contingent outcome distributions that are closest to the one induced by action $a$. Here the “closest” distributions are those in the support of the agent’s prior belief that minimize the Kullback-Leibler divergence from the true distribution given that the agent plays $a$; we call these the “KL-minimizers for $a$.” The stronger concept of uniform Berk-Nash equilibrium requires that the action is a best response to any mixture over the KL minimizers, as opposed to at least one of them.

We then investigate sufficient conditions for two alternative definitions of what it means for an action to be a long-run outcome. We say that an action is stable if play converges to it with high probability for some open set of initial beliefs. Proposition 2 shows that every uniformly strict Berk-Nash equilibrium is stable, regardless of the agent’s discount factor, where “strict” indicates that the action is the strict myopic best response to the agent’s beliefs, and “uniformly” requires that this is true for all of the closest outcome distributions (as opposed to being true for at least one of them). The converse is not true in general, but Proposition 3 shows that under a minor technical condition the only stable equilibria are the uniform Berk-Nash equilibria.

We say that an action is positively attracting if there is positive probability that it is the limit outcome under every optimal policy for every full-support prior belief. Our setup allows us to model a number of different forms of misspecified learning. In particular, in the case of
causation neglect, where the agent mistakenly believes that the distribution of outcomes is the same for all actions, and in subjective bandit problems, where the agent mistakenly believes that the outcomes observed when playing one action are uninformative about the outcome distributions induced by others, we obtain partial converses to Proposition 1: All uniformly strict equilibria are positively attractive, meaning that they have positive probability of being the limit outcome from any starting belief. Moreover, in subjective bandit problems that are weakly identified (Esponda and Pouzo (2016)) we can relax uniformly strict to strict.

To prove these results, we first extend Diaconis and Freedman (1990)’s result that Bayesian updating is uniformly consistent to the case of misspecified prior beliefs, a fact that may be of use in future work. We use this extension to guarantee that the agent starts to play the equilibrium action with positive probability. Then, we use the stability result of Proposition 2 to show that with positive probability the agent uses the action forever. We also observe that in a supermodular decision problem, extreme uniformly strict equilibria are positively attractive. There the additional structure of the problem lets us dispense with the first step of the proof.

We also generalize our results to a setting in which the agent observes a signal before taking an action. Here too a limit action must be a Berk-Nash equilibrium. Moreover, if the agents ignore the predictive value of the signals, a behavior that we label signal neglect, every uniformly strict Berk-Nash equilibrium is positively attractive.

We illustrate how the sorts of misspecification that can be modeled in our environment correspond to incorrect beliefs about the extensive form of a game, and also with a running example of a central bank choosing an exchange-rate policy. The effectiveness of its policy depends on the strength of the economy, summarized by the probability of an expansion in each period and the probability of a speculative attack on the domestic currency. One form of misspecification is that the central bank may not assign positive probability to the correct map from the policy adopted to the probability distribution over outcomes. A second is that the central bank may have causation neglect and think that its policy does not influence the probability distribution over outcomes.

1.1 Related Work

Here we briefly discuss the most relevant prior work. We make more detailed comments about the links to the literature and how our results compare to previous ones after introducing the formal model.

The statistics literature starting with Berk (1966) studies misspecified Bayesian learning
in environments where the distribution (or stochastic process) of outcomes is exogenous. In many economic applications, actions and associated signal distributions aren’t fixed but change endogenously over time depending on an action taken by the agent (Arrow and Green [1973], Nyarko [1991], Esponda [2008]). Esponda and Pouzo (2016) 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 their observations on the support of their prior. They also analyze dynamics when the payoff function is subject to small random shocks.

Heidhues, Koszegi, and Strack (2018) assume that the agent’s subjective belief is updated according to a linear updating rule. Frick, Iijima, and Ishii (2019b) assume that the agent’s prior is concentrated on a finite set of states, and they make a continuity assumption that our model does not satisfy. Esponda, Pouzo, and Yamamoto (2019) use stochastic approximation techniques to study when the action frequencies converge, as opposed to our focus on whether play converges to a fixed action. Fudenberg, Romanyuk, and Strack (2017) give a complete characterization of limit actions and beliefs for possibly non-myopic learners in a binary prior model with Brownian noise in continuous time. Molavi (2019) studies misspecification in a temporary equilibrium model of macroeconomics; his leading example is where agents mistakenly think that some variables have no impact.¹

All the previous papers consider misspecified Bayesian agents. There is also a literature that studies the long-run outcomes under different learning heuristics. Such heuristics are due to misspecification in the sense that the agent is unable to formulate a probabilistic assessment of the data generating process. Many of these heuristic features some form of neglect of relevant elements of the environment similar to the ones we consider in our Section 4 (see, e.g., Tversky and Kahneman, [1973], Rabin and Schrag, [1999] and Jehiel, 2018).

¹Other recent papers studying misspecification in specific examples of single agent decision problems with endogenous data are Gagnon-Bartsch, Rabin, and Schwartzstein (2018), He (2019), Heidhues, Kőszi, and Strack (2018). Misspecification has also been studied in settings of social learning, where agents learn from observations generated by others as in Bohren (2016), and Bohren and Hauser (2018), and Frick, Iijima, and Ishii (2019a).
2 The Model

2.1 Setup

Actions, Utilities and Objective Outcome Distributions  We consider a discrete time problem: In each period \( t \in \{1, 2, 3, \ldots \} \) an agent chooses an action from the finite set \( A \).\(^2\) This choice has two effects. First, each action \( a \in A \) induces an objective probability distribution \( p^*_a \in \Delta(Y) \subset \mathbb{R}^{|Y|} \) over the finite set of possible outcomes \( Y \).\(^3\) Second, the action, paired with the realized outcome, determines the flow payoff of the agent via the utility function \( u: A \times Y \rightarrow \mathbb{R} \).

Formally, we consider the probability space \( (\Omega, \mathcal{F}, \mathbb{P}) \). The sample space \( \Omega = (Y^\infty)^A \) consists of infinite sequences of action dependent outcome realizations \( (x_{a,1}, x_{a,2}, \ldots)_{a \in A} \), and \( x_{a,k} \) determines the outcome when the agent takes the action \( a \) for the \( k \)-th time. \( \mathcal{F} \) is the product sigma algebra and the probability measure \( \mathbb{P} \) is the product measure induced by independent draws from the relevant component of \( p^* \). We denote the outcome observed by the agent in period \( t \) after action \( a_t \) by \( y_t = x_{a_t,k} \), where \( k = |\{\tau \leq t: a_\tau = a_t\}| \) is the number of times the agent has taken action \( a_t \) up to and including period \( t \).

Subjective Beliefs of the Agent  The agent correctly believes that the map from actions to probability distributions over outcomes is fixed and depends only on their current action, but they are uncertain about the distribution each action induces. Let \( P = \times_{a \in A} \Delta(Y) \subset \mathbb{R}^{|Y| \times |A|} \) be the space of all action-dependent outcome distributions, and let \( p_a \in \Delta(Y) \) denote the \( a \)-th component of \( p \in P \). We endow \( P \) with the sup-norm topology. The agent’s uncertainty is captured by a prior belief \( \mu_0 \in \Delta(P) \), where \( \Delta(P) \) denotes the metric space of Borel probability measures on \( P \) endowed with the Prokhorov metric, so that it has the topology of weak convergence of measures. Here the support of \( \mu_0 \) is the set of distributions over outcomes that the agent thinks are possible. We call these the conceivable outcome distributions, and denote them as \( \Theta = \text{supp} \mu_0 \). We do not require that the agent’s model is correctly specified, i.e. that the true outcome distribution \( p^* \) is conceivable. The only restriction that we impose is that if \( p \) is conceivable, then no finite number of observations can eliminate it from the support. Formally, we will maintain the following assumption:

Assumption 1. For all \( p \in \Theta \) and \( \varepsilon > 0 \), there exists \( p' \in \Theta \) with \( ||p' - p|| < \varepsilon \) such that for all \( a \in A \), if \( p^*_a(y) > 0 \) then \( p'_a(y) > 0 \).

\(^2\)We endow \( A \), as well as any other finite set, with the discrete topology.
\(^3\)Throughout we denote objective distributions with a superscript \( * \).
Our specification allows the agent’s subjective uncertainty to be correlated across actions. For example, under causation neglect, the agent has a belief about action-contingent distributions that is perfectly correlated: they are certain that every action generates the same outcome distribution.

**Updating of Subjective Beliefs** We assume throughout that the agent updates their beliefs using Bayes rule. Given a sequence of realized actions and outcomes \((a^t, y^t)\), let

\[
L_t(a^t, y^t, p) = \prod_{\tau=1}^{t} p_{a_{\tau}}(y_{\tau})
\]

(1)
denote the likelihood of observing \((a^t, y^t)\) up to period \(t\) when the distribution over outcomes equals \(p\).

Denote by \(\mu_t(\cdot | (a^t, y^t))\) the subjective belief the agent obtains using Bayes rule after action sequence \(a^t\) and signal sequence \(y^t\). Since the agent’s prior has support \(\Theta\), any posterior belief \(\mu_t(\cdot | (a^t, y^t))\) lies in the set \(\Delta(\Theta) := \mathcal{M}\). Moreover from Bayes rule we have that for every \(\Psi \subseteq \Theta\)

\[
\mu_t(\Psi | (a^t, y^t)) = \frac{\int_{\Psi} L_t(a^t, y^t, p)d\mu_0(p)}{\int L_t(a^t, y^t, p)d\mu_0(p)}. 
\]

(Bayes Rule)

We sometimes suppress the dependence of the posterior belief on the realized sequence and just write \(\mu_t\).\(^4\)

**Behavior of the Agent** A (pure) policy \(\pi : \bigcup_{t=0}^{\infty} A^t \times Y^t \rightarrow A\) specifies an action for every history \((a_1, y_1, a_2, y_2, \ldots, a_t, y_t)\), and an initial action \(a_1\). We assume that the agent’s objective is to maximize the expected discounted value of per-period utility with discount factor \(\beta \in [0,1)\), and restrict to optimal policies. Throughout, we denote by \(a_{t+1} = \pi(a^t, y^t)\) the action taken in period \(t\), where \((a^t, y^t) = (a_s, y_s)_{s=1}^t\) is the sequence of realized actions and outcomes. Together, the probability measure \(\mathbb{P}\) and a policy \(\pi\) induce a probability measure \(\mathbb{P}_\pi\) on \((a_\tau, y_\tau)_{\tau=1}^{\infty}\).\(^5\)

Standard results guarantee that in this setting there is an optimal policy \(\pi\) that depends

---

\(^4\) Assumption \(^1\) guarantees that Bayesian updating is always well defined.

\(^5\) Multiple state spaces lead to the same law for the stochastic processes we are interested in. In particular, we could have started from the probability space of action-dependent outcome realizations \((x_{a,1}, x_{a,2}, \ldots)_{a \in A}\) but with \(x_{a,k}\) denoting the outcome realization if the agent takes action \(a\) at period \(k\). An argument similar to that of Lemma 5 of Fudenberg and He (2017) shows that this choice would not affect our results.
on the history only through the agent’s beliefs, that is,

\[ \mu_t(\cdot \mid (a^t, y^t)) = \mu_{t'}(\cdot \mid (a^{t'}, y^{t'})) \implies \pi(a^t, y^t) = \pi(a^{t'}, y^{t'}). \]

In what follows, we will maintain the assumption that the policy of the agent satisfies this restriction.

For every belief \( \nu \in \Delta(P) \) we denote by \( \nu_a \) the marginal distribution associated with action \( a \), i.e.,

\[ \nu_a(\Psi) = \int p_{a \in \Psi} d\nu (p) \text{ for all } \Psi \subseteq \Delta(\Delta(Y)). \]

We denote by \( \mathbb{E}_{p_a} [f(y)] = \sum_{y \in Y} f(y)p_a(y) \) the expectation of \( f : Y \to \mathbb{R} \) under the outcome distribution \( p_a \in \Delta(Y) \). \( A^m(\nu) \) denotes the set of myopic best replies to the belief \( \nu \), i.e.,

\[ A^m(\nu) = \operatorname{argmax}_{a \in A} \int_{\Delta(Y)} \mathbb{E}_{p_a}[u(a, y)] d\nu_a(p_a). \] (2)

We say that the agent acts myopically if for all \((a^t, y^t) \in \bigcup_{\tau=0}^{\tau} A^\tau \times Y^\tau \) we have \( \pi(a^t, y^t) \in A^m(\mu_t(\cdot \mid a^t, y^t)) \).

### 2.2 Forms of Misspecification

Our model encompasses many sorts of misspecified learning. Here we summarize the cases we analyze. We give detailed economic examples of these biases and their implications in Section 3.

#### 2.2.1 Causation Neglect

An agent with causation neglect believes that the realized outcome is not affected by the chosen action. This corresponds to the case in which the prior assigns positive probability only to distributions over outcomes that are independent of the chosen action.

**Definition 1** (Causation Neglect). We say that an agent has causation neglect if for every \( a, a' \in A \) and every \( p \in \Theta \), we have \( p_a = p_{a'} \).

Note that this definition does not say whether the action really does influence the distribution or not; if it does not then it is not a mistake to “neglect” it.

An agent with causation neglect can be misspecified in two ways: Their belief that the action doesn’t change the distribution of outcomes might be incorrect, and the support of their prior belief might not include the true distribution over outcomes for some actions (i.e. there exists \( a \in A \) with \( p_a^* \notin \{p_a \in \Delta(Y) : p \in \Theta\} \)).
An agent with causation neglect updates their belief about the outcome distribution as if they faced an i.i.d. environment. This allows us to use a novel extension of the Diaconis and Freedman (1990) result about uniform consistency with misspecified beliefs to guarantee that the beliefs will concentrate on the conceivable outcome distributions closest to the empirical average. We use this result show that if \( a \) is a uniformly strict Berk-Nash equilibrium, there is a sequence of outcomes such that the best reply to the induced belief is \( a \), and such that with positive probability the agent sticks to action \( a \) forever. This means that under causation neglect, all uniformly strict Berk-Nash equilibria are positively attractive. Notice that under correlation neglect generically all the Berk-Nash equilibria are uniformly strict.

2.2.2 Subjective Bandit Problems

The other extreme case encompassed by our setup is where the agent thinks that they face a bandit problem, that is that the probability distributions over outcomes induced by different actions are subjectively independent. This corresponds to the case where the agent’s prior \( \mu_0 \) is a product measure.

**Definition 2** (Bandit Problem). We say that an agent faces a subjective bandit problem if \( \mu_0 = \times_{a \in A} \mu_{0,a} \in (\Delta (\Delta (Y)))^A \). Each \( \mu_{0,a} \in \Delta (\Delta (Y)) \) is the agent’s prior about the distribution over outcomes induced by action \( a \).

We use our extension of Diaconis and Freedman (1990) to show that uniformly strict Berk-Nash equilibria are positively attractive in this setting as well, provided that the agent is sufficiently patient.\(^6\)

2.2.3 One Dimensional Parametric Decision Problems

In one-dimensional parametric decision problems, the agent’s uncertainty is summarized by a parameter \( \gamma \in \mathbb{R} \). This restriction makes it easier to rule out cycles and guarantee convergence. The parameter determines the distribution over outcomes through a function \( \phi(\gamma) \) which maps parameters to action-dependent outcome distributions.

**Definition 3** (One-Dimensional Parametric Decision Problems). The decision problem is one-dimensional parametric if there exists \( \Gamma \subseteq \mathbb{R} \) and a function \( \phi : \Gamma \rightarrow P \) such that

\(^6\)The proof shows that if \( b \) is a uniformly strict Berk-Nash equilibrium and the agent is very patient, then there is positive probability that the agent’s beliefs eventually give \( b \) the highest Gittins index. Note that the agent’s discount factor is irrelevant under causation neglect since there the agent does not think there is any information value in experimenting with other actions.
A one-dimensional decision problem is *supermodular* if the actions $A$ can be ordered such that $(\gamma, a) \mapsto \mathbb{E}_{\phi(\gamma)}[u(a, y)]$ is supermodular.\footnote{Formally, if $a > a'$ and $\gamma > \gamma'$ then $\mathbb{E}_{\phi(\gamma)}[u(a, y)] + \mathbb{E}_{\phi(\gamma')}[u(a', y)] \geq \mathbb{E}_{\phi(\gamma)}[u(a, y)] + \mathbb{E}_{\phi(\gamma')}[u(a', y)]$.}

Esponda, Pouzo, and Yamamoto (2019) consider one-dimensional decision problems in which for every action $a$ the set of action-contingent outcome distributions that minimize the Kullback-Leibler divergence is a singleton, and the true data generating process corresponds to a value of $\gamma$ that may be outside the support of the agent’s prior. They show that the action process converges almost surely. They also show that, if the decision problem is supermodular and all pure Berk-Nash equilibria are uniformly strict, play converges to one of the Berk-Nash equilibria.

Heidhues, Kőszegi, and Strack (2018) show that in supermodular decision problems where the outcomes are real numbers and $\phi$ is an additive shift, whenever there is a unique Berk-Nash equilibrium play converges there from any starting belief. Our Example 3 shows that their result does not hold in our more general structure: the unique uniformly strict Berk-Nash equilibrium may not be positively attractive. We show that when a stronger version of supermodularity is imposed, our positive attractiveness results can extend to extremal uniformly strict Berk-Nash equilibria.

### 2.2.4 Binary Prior

Another assumption that is commonly made to facilitate the analysis of misspecified models is that of a binary prior, that is $|\Theta| = 2$. Fudenberg, Romanyuk, and Strack (2017) characterizes the long-run behaviour in the case of a binary prior when the outcome is the sum of the chosen action and a Brownian motion. Bohren (2016) and Bohren and Hauser (2018) analyze misspecified binary-prior models in the context of social learning.

### 2.2.5 Finite Support

A final assumption that is commonly made when studying misspecified learning is that the support of the prior is finite. Our general setup encompasses this case as well, which allows us to highlight an important difference between environments with finite or infinite support. With a finite-support prior, if behavior converges to an action $a$, $a$ is a *uniform* Berk-Nash equilibrium. That is, $a$ is a best reply to all the outcome distributions that minimize the Kullback-Leibler divergence from $p^*_a$. However, we show by example that non uniform Berk-Nash equilibria can be limit points when the support of the prior is infinite.
Frick, Iijima, and Ishii (2019b) assumes a finite prior, but our setup does not nest theirs as it does not satisfy their continuity assumption, which implicitly requires a continuum of actions. However, we share their interest in the stability properties of equilibria.

2.2.6 Signal Neglect

Here we suppose that each period the agent observes a finite-dimensional signal vector $s \in S$ before taking an action $a \in A$. The signal may convey information about the outcome distribution, and it may also directly enter the payoff function.

We allow the agent to be uncertain about outcome distributions induced by various signals and actions. Let $P = (\Delta(Y))^{|A| \times |S|} \subset \mathbb{R}^{|Y| \times |A| \times |S|}$ be the space of all signal and action dependent outcome distributions. The agent’s belief is a probability measure $\mu$ over $P$, where $p_{s,a}(y)$ denotes the probability under $p$ of outcome $y$ after observing signal $s$ and playing action $a$.

Definition 4 (Signal Neglect). An agent has signal neglect if for every $s, s' \in S$ and every $a \in A$, and $p \in \Theta$, we have $p_{s,a} = p_{s',a}$.

As an example, consider a seller who chooses whether to undertake an investment (e.g., hiring an additional worker for the day) after observing the number of people in the market that day. The agent realizes that the signal is payoff relevant, but they do not realize that when fewer consumers show up, a higher fraction of them buy. We show that this form of misspecification will result in underinvestment when market attendance is high. We discuss this example in more detail in Section 6.

3 Limit Points and Berk-Nash Equilibria

We will be interested in the possible limit points of the action process, and when it converges to one of them. Note that these are different questions than whether the agent’s beliefs converge: Beliefs can oscillate when actions are fixed, as in Berk’s example where there the agent doesn’t have an action choice, and conversely actions can oscillate with fixed beliefs if the agent is indifferent. The possibility that beliefs oscillate with fixed actions is the driving force behind the uniformity requirement in several of our results, see e.g., Proposition (ii)

---

8 More generally the agent might mistakenly think that some components of the signal vector have no effect on outcomes, as in Molavi (2019)

9 The possibility that beliefs oscillate with fixed actions is the driving force behind the uniformity requirement in several of our results, see e.g., Proposition (ii)
log-likelihood ratio between these two states is a random walk and thus oscillates between assigning high probability to each of the two states. Conversely, if the agent is indifferent between multiple actions at the limit belief, their actions might not converge even though their beliefs do.

Formally, the action process converges to action $a$ if there exists a time period $T \in \mathbb{N}$ such that $a_t = a$ for all later time periods $t > T$. We say that the action process converges to $a$ with positive probability (resp. with probability 1) under policy $\pi$ if there is a measurable set $C \subseteq A^\infty \times Y^\infty$ with $\mathbb{P}_\pi[C] > 0$ (resp. with $\mathbb{P}_\pi[C] = 1$) such that $a_t$ converges to $a$ in $C$. Note that a given prior may be consistent with multiple optimal policies, and which policy is used can influence whether the action process converges and if so to which points.

The concept of Berk-Nash Equilibria (Esponda and Pouzo, 2016) will play a key role in our analysis. Intuitively, a Berk-Nash equilibrium is an action $a$ such that there exists a belief that makes the action optimal, and which gives positive probability only to the conceivable outcome distributions that best match the objective outcome distribution $p^*_a$. Formally, given two distributions over outcomes $q, q' \in \Delta(Y)$ we define

$$H(q, q') = -\sum_{y \in Y} q(y) \log q'(y).$$

Recall that for every $p$, $p_a$ denotes the outcome distribution corresponding to action $a$. We let $\tilde{P}_a$ denote the set of conceivable outcome distributions for action $a$ that best match $p^*_a$:

$$\tilde{P}_a = \text{argmin}_{p_c : p \in \Theta} H(p^*_a, p_a) \subset \Delta(Y).$$

And for each action $a$, let

$$D_a = \{ p \in \Theta : p_a \in \tilde{P}_a \} = \text{argmin}_{p \in \Theta} H(p^*_a, p_a) \subset P$$

(3)

denote the set of action-contingent outcome distributions that minimize the Kullback-Leibler divergence between distributions in the support of $\mu_0$ and the true distribution $p^*_a$ given that $a$ is played.\textsuperscript{10} Note that the elements of $D_a$ are elements of $P$, and so specify an outcome distribution for each action $a' \in A$, even though which elements of $P$ are included only depends on the distribution $p^*_a$ induced by $a$ and on the outcome distributions that $\mu$ thinks $a$ might induce. We call this the set of “KL-minimizers for action $a$ and belief $\mu$."

\textsuperscript{10}The Kullback-Leibler divergence between $p^*_a$ and $p_a$ is given by $-\sum_{y \in Y} p^*_a(y) \log(p^*_a(y)) + H(p^*_a, p_a)$ so any $p_a$ that minimizes $H(p^*_a, p)$ also minimizes the Kullback-Leibler divergence between $p^*_a$ and $p_a$.\textsuperscript{10}
Note also that if \( p^* \in \Theta \) then each minimizing \( p \) has \( p_a = p^*_a \). In particular this is true if \( \mu \) has full support.

Following Berk (1966), we expect the agent’s beliefs to concentrate on the KL minimizers for action \( a \) and belief \( \mu_0 \) if an agent with prior belief \( \mu_0 \) always plays \( a \). This motivates Esponda and Pouzo (2016)’s notion of a Berk-Nash equilibrium. We will use several variations of this concept to capture various senses in which an action is or is not a long-run outcome of the agent’s learning process.

**Definition 5 (Berk-Nash Equilibrium).**

(i) Action \( a \in A \) is a **Berk-Nash equilibrium** if there exists a belief \( \nu \in \Delta(D_a) \) such that \( a \) is myopically optimal given \( \nu \), i.e. \( a \in A^m(\nu) \).

(ii) Action \( a \) is a **strict Berk-Nash equilibrium** if for some belief in \( \nu \in \Delta(D_a) \), \( a = A^m(\nu) \), i.e. \( a \) is the unique myopically optimal action.

(iii) Action \( a \) is a **uniform Berk-Nash equilibrium** if for all \( q \in \bar{P}_a \) there exists a \( \nu \in \Delta \{ p \in D_a : p_a = q \} \) such that \( a \in A^m(\nu) \).

(iv) Action \( a \) is a **uniformly strict Berk-Nash equilibrium** if \( \{ a \} = A^m(\nu) \) for every belief in \( \nu \in \Delta(D_a) \).\(^{11}\)

The intuition behind the definition of uniform Berk-Nash equilibrium is the following. Playing action \( a \) induces distribution \( p^*_a \) over the observed outcomes, and the set \( \bar{P}_a \) includes all the outcome distributions that best match \( p^*_a \). Uniformity requires that for each \( p_a \in \bar{P}_a \) there exists a belief about the outcome distributions induced by the other actions that justifies \( a \) as a best reply. Observe that the difference between Berk-Nash equilibria and uniform Berk-Nash equilibria is peculiar to the misspecified case. If the agent is correctly specified, both of them coincide with self-confirming equilibrium.

The following result motivates our focus on (uniform) Berk-Nash equilibria.

**Proposition 1 (Limit Actions are Berk-Nash Equilibria).** Suppose the agent uses an optimal myopic policy.

(i) If actions converge to \( a \in A \) with positive probability, then \( a \) is a Berk-Nash equilibrium;

(ii) If \( \Theta \) is finite and actions converge to \( a \in A \) with positive probability, then \( a \) is a uniform Berk-Nash equilibrium.

The first part of this proposition follows from the more general Proposition\(^8\) whose proof is in the Appendix. In outline, the result follows from the fact that if actions converge to \( a \)

\(^{11}\)Esponda, Pouzo, and Yamamoto (2019) call this a strict equilibrium.
then after some time the agent plays that action in all future periods, and Berk (1966)’s result
that the agent’s beliefs converge to the set of KL minimizers when their observations are a
sequence of i.i.d. signals. Example 5 shows that this conclusion fails without Assumption 1.

The second part of the proposition assumes that the set of KL minimizers contains only
a finite number of points. The assumption implies that the agent’s subjective log-likelihood
ratio is a finite-dimensional martingale. The proof combines the Central Limit Theorem
with the Kochen-Stone Lemma to prove that for every \( p_a \in \bar{P} \), the posterior will eventually
assign an arbitrarily high probability to the models \( \hat{p} \in \Theta \) that agree with \( p_a \) conditional on
action \( a \), i.e., such that \( \hat{p}_a = p_a \). But then, unless \( a \) is a uniform Berk-Nash equilibrium,
it will eventually no longer be played. Example 7 in the Appendix shows that in general we
cannot dispense with the assumption of a finite support.

**Remark 1.** We do not know the exact conditions that ensure actions can converge only to
uniform Berk-Nash equilibria. However, we conjecture that a sufficient condition is that the
density of the prior around two different minimizing distributions is bounded.

Similar results to the first part of the proposition were obtained in related settings by
Esponda and Pouzo (2016), Esponda, Pouzo, and Yamamoto (2019), and Frick, Iijima, and
Ishii (2019b). The first part of our proposition is most similar to Proposition 1 of Esponda,
Pouzo, and Yamamoto (2019), and is a consequence of it modulo some technical conditions.\(^1\)

We do not know of any result that is analogous to the second part of Proposition 1,
which establishes conditions under which the optimality of the action at all KL-minimizers
is necessary for convergence and thus suggests some type of uniformity condition should
be imposed as a refinement of Berk-Nash equilibrium. Indeed, the only other equilibrium
refinement we know of that tests for optimality against all beliefs in a non-singleton set is
Fudenberg and He (2020) which also has a foundations in learning theory.\(^2\)

### 4 Sufficient Conditions for Long-Run Persistence

Proposition 1 shows that play can only converge to a given action \( a \) if that action is a Berk-
Nash equilibrium. This section gives sufficient conditions for two different senses of what it

\(^{12}\)Thus the argument extends Berk (1966)’s example where two different conceivable probability distributions
are the best explanations of the observed data, and the log-likelihood ratio is a random walk.

\(^{13}\)In particular, our Assumption \( \# \) is weaker than our Assumption 2 (iii), although our Proposition 8 is
not. Esponda and Pouzo (2016)’s similar result assumed that the utility function is subject to i.i.d. shocks
as in Fudenberg and Kreps (1993).

\(^{14}\)However, Fudenberg and He (2020) studies a model of correctly specified learning, and does not study
explicit non-equilibrium dynamics.
means for an action to be a plausible long run outcome.

4.1 Stability

We say that action \( a \) is “stable” if play is very likely to converge to \( a \) starting from some beliefs in a neighborhood of a KL-minimizer for \( a \). For \( \nu \in \mathcal{M} \), let \( B_\varepsilon(\nu) = \{ \nu' \in \mathcal{M} | d(\nu', \nu) \leq \varepsilon \} \) be the set of conceivable distributions that are within \( \varepsilon \) of \( \nu \).

**Definition 6 (Stability).**

(i) A Berk-Nash equilibrium \( a \) is *stable* for discount factor \( \beta \) if there is a belief \( \nu \in \Delta(D_a) \) such that for every \( \kappa \in (0, 1) \), there exists an \( \varepsilon > 0 \) such that for all prior beliefs in \( B_\varepsilon(\nu) \) the action prescribed by any optimal policy converge to \( a \) with probability larger than \( (1 - \kappa) \).

(ii) A Berk-Nash equilibrium \( a \) is uniformly stable for discount factor \( \beta \) if for every \( \kappa \in (0, 1) \), there is an \( \varepsilon > 0 \) such that for all prior beliefs \( \nu \in \mathcal{M} \) such that \( \nu(D_a) > 1 - \varepsilon \) the action prescribed by any optimal policy converge to \( a \in A \) with probability greater than \( 1 - \kappa \).

Once again, the uniformity requirement has bite only if the agent is misspecified: if \( p^* \in \Theta \), the actions that are stable are uniformly stable, and they coincide with strict self-confirming equilibria, or when there is more than one KL minimizer.

The next proposition shows that if a Berk-Nash equilibrium is uniformly strict then convergence to that equilibrium is arbitrarily likely for some beliefs:

**Proposition 2.** Every uniformly strict Berk-Nash equilibrium is uniformly stable for any discount factor \( \beta \).

Frick, Iijima, and Ishii (2019b) obtain a similar conclusion in a model where the agent’s prior has finite support. This conclusion is sharper than showing there are beliefs under which there is positive probability of converging to the equilibrium, as in Proposition 4 of Esponda, Pouzo, and Yamamoto (2019).

The proof of Proposition 2 first derives a neighborhood of action-dependent outcome distributions that are close to the Kullback-Leibler minimizers such that if the beliefs assign a sufficiently high probability to that neighborhood, the optimal action is the uniformly strict Berk-Nash equilibrium \( a \). That such a neighborhood exists for a myopic policy follows from the definition of uniformly strict Berk-Nash equilibrium \( a \). Under a non-myopic policy, since beliefs are not degenerate, some actions may have an experimentation value. However, when
the beliefs are sufficiently concentrated around the minimizers, the value of any alternative action cannot be much higher than its value against the most favorable minimizer, and since $a$ is a uniformly strict Berk-Nash equilibrium, this value is strictly lower than that of $a$. Then we use the continuity of the Renyi divergence at $l = 0$, as in Frick, Iijima, and Ishii (2019b), together with a generalization of the arguments in Fudenberg and Levine (1992) to the misspecified case, to argue that a transformation of the odds ratio of this neighborhood is a positive supermartingale, under the outcome distribution induced by action $a$. Finally, we use the Dubins’ inequality to guarantee that if the probability initially assigned to the neighborhood is sufficiently high, that probability never crosses the threshold that makes action $a$ suboptimal.

Proposition 2 is in contrast to Heidhues, Koszegi, and Strack (2018) who find that only some uniformly strict Berk-Nash equilibria are stable in a model with a continuum of actions and a convex, real-valued set of states. We believe that the key reason for the different conclusion is that in their model actions that are sufficiently near the strictly best response incur arbitrarily small losses, and are best responses to nearby beliefs.

The next example shows that Proposition 2 does not extend to strict Berk-Nash equilibria that are not uniformly strict.

**Example 1.** [A strict but not stable Berk-Nash equilibrium] Suppose there are two actions $a$ and $b$, and two outcomes $Y = \{0, 1\}$. We identify the elements of $\Delta(Y)$ with the probability they assign to outcome 1. Let $p^*_a = \frac{1}{2}$, and suppose that the agent believes that the outcome distribution does not depend on the action, so they have correlation neglect, and let $\Theta = \left[0, \frac{1}{4}\right] \cup \left[\frac{3}{4}, 1\right]$. Suppose that $\Delta^{m} (\delta_{3/4}) = \{a\}$, and $\Delta^{m} (\delta_{1/4}) = \{b\}$.

Here $a$ is a strict Berk-Nash equilibrium supported by the belief $\delta_{3/4}$, but it is not uniformly strict. Now let $\nu_\epsilon = \epsilon \delta_{1/4} + (1 - \epsilon) \text{U}\left[\frac{3}{4} + \epsilon, \frac{3}{4} + 2\epsilon\right]$, and note that $\arg\min_{\text{supp}\nu_\epsilon} H (p^*_a, p_a) = \left\{\frac{1}{2}\right\}$ and $a \notin \Delta^{m} (\delta_{1/4})$. Note also that as $\epsilon$ goes to 0, $\nu_\epsilon$ converges to $\delta_{3/4}$. Yet by Proposition 1 the probability of converging to a starting from belief $\nu_\epsilon$ is 0 for all $\epsilon$.

**Example 2** (A Stable Berk-Nash Equilibrium that is not uniformly strict). To see that the converse of Proposition 2 is false, suppose there are 2 actions $a$ and $b$ that induce the same distribution on $Y = \{0, 1\}$. Suppose also that $a$ is the unique best response to any belief has positive probability on 1, while both $a$ and $b$ are best responses when sure the outcome will be 0. Here $a$ is not uniformly strict but it is stable.

The next result is a partial converse to Proposition 2 and it shows a key difference between misspecified and correctly specified environments. In misspecified settings, more than one
outcome distribution may be the best explanation of the outcome distribution generated by an action \( a \), and if \( a \) is not a uniform equilibrium, there is positive probability that the agent’s belief will eventually concentrate on outcome distributions that induce a different best reply \( a' \neq a \).

**Proposition 3.** Suppose that \( a \) is not a uniform Berk-Nash equilibrium. If for every \( p \in D_a \) and \( \varepsilon > 0 \) there exists \( q \in B_\varepsilon(p) \), \( q \notin D_a \), then \( a \) is not stable under a myopic policy.

If \( a \) is not a Berk-Nash equilibrium, the result immediately follows from Proposition 1. When \( a \) is a Berk-Nash equilibrium but not a uniform Berk-Nash equilibrium, there is at least one \( p \in D_a \) for which the best reply is not \( a \). Suppose that the agent’s initial belief \( \nu \) assigns a very low positive probability to this \( p \), and the remaining probability to distributions \( p' \) that are close to distributions in \( D_a \) but are not in \( D_a \). Then, \( \{p\} = \text{argmin}_{p' \in \supp \nu} H(p^*_a, p'_a) \), and Proposition 1 implies that since \( a \notin A(\delta_p) \), the probability of converging to \( a \) starting with belief \( \nu \) is zero.

### A Characterization of Berk-Nash Equilibria
Combining the results of this section with Proposition 1 gives a sharp relationship between equilibria and stable outcomes.

**Corollary 1.** If \( |\Theta| < \infty \) and \( \beta = 0 \), the following are equivalent:

(i) \( a \in A \) is a uniform Berk-Nash equilibrium.

(ii) There exists a belief \( \nu \in \mathcal{M} \) with \( D_a \subseteq \supp \nu \) and an optimal policy such that actions converge to \( a \in A \) with positive probability.

**Corollary 2.** For every discount factor \( \beta \in [0, 1) \) the following are equivalent:

(i) \( a \in A \) is a uniformly strict Berk-Nash equilibrium.

(ii) \( a \in A \) is uniformly stable for \( \beta \).

Theorem 1 of Frick, Iijima, and Ishii (2019b) characterizes the locally stable beliefs for finite-support priors. Because the associated actions and preferences are left implicit it is difficult to compare their results with ours, but as we pointed out earlier, actions can converge even when beliefs do not.

---

\(^{15}\)We could have defined stability with respect to a myopic policy instead of all of them. Under the assumptions of the next proposition, actions that are not uniform Berk-Nash equilibria would fail the stability test even under that more permissive definition.

\(^{16}\)The hypothesis of Proposition 1 excludes finitely supported priors, but in this case Proposition 1 guarantees that the probability of converging to a non-uniform equilibrium is 0. A fortiori, this means that non-uniform equilibria are not stable.
4.2 Positive Attractiveness

The previous section gave sufficient conditions for an action to be played in the long-run with high probability for some prior beliefs. Another natural notion of a being a long-run outcome is that for every prior belief there is strictly positive probability that the agent’s action converges to $a$.

We say that an action is positively attracting if there is positive probability that it is the limit outcome under every optimal policy.

**Definition 7 (Positively Attracting).** The action $a \in A$ is positively attracting if for every optimal policy $\pi$

$$\mathbb{P}_\pi \left( \lim_{t \to \infty} a_t = a \right) > 0.$$ 

In later sections we give sufficient conditions for uniformly strict Berk-Nash equilibria to be positively attracting. Benaïm and Hirsch (1999) prove a similar result for the linearly stable Nash equilibria of stochastic fictitious play, which is based on correctly-specified Bayesian learning, but positive attractiveness has not previously been studied in the mis-specified context.

These arguments rely on Lemma 4 in the appendix, which shows that beliefs about the outcome distribution concentrate around the distributions that best fit the empirical frequency of outcomes. Importantly, this result does not require that either actions or empirical frequencies converge. It is based on arguments made in Diaconis and Freedman (1990), who considered agents with full support beliefs. It will be important in what follows that these results apply pathwise.

4.2.1 Causation Neglect

When the agent has causation neglect they believe that the distribution over outcomes is the same for all actions. Proposition 1 gives a necessary condition for the convergence of beliefs and actions. Example 3 shows that this condition is not sufficient to ensure convergence, even when there is a unique Berk-Nash equilibrium, and this equilibrium is uniformly strict.

**Example 3.** In this example the agent has three actions and their prior has support $\{p^1, p^2, p^3\}$. The details are in the following table.
In this example $a = 3$ is the only Berk-Nash equilibrium and is uniformly strict. However, if the agent takes an action $a \in \{1, 2\}$ then the subjective likelihood assigned to $p^3$ goes down and thus play never converges to $a = 3$ if the prior assigns sufficiently low probability to $p^3$.

The next proposition gives a sufficient condition for a Berk-Nash equilibrium to be positively attracting.

**Proposition 4.** Suppose that the agent has causation neglect. If $a$ is a uniformly strict Berk-Nash equilibrium such that $p^*_{a}$ is absolutely continuous with respect to $p^*_{a'}$ for all $a' \in A$, then it is positively attracting.

The proposition’s assumption of causation neglect guarantees that the uncontingent empirical distribution is a sufficient statistic for the agent’s beliefs. To prove the result, we first use Lemma 4 to show that on every path of outcome realizations, these beliefs concentrate around the distributions that minimize the KL divergence from the empirical frequency. We use this concentration to show there is a finite sequence of outcomes that has positive probability and leads the agent to play $a$. Since $a$ is a uniformly strict Berk-Nash equilibrium, if beliefs concentrate around the minimizers, $a$ becomes the unique best reply. While using $a$, the relative probability assigned to distributions in $D_a$ increases in expectation so we can combine Dubins’ upcrossing inequality with the fact that $a$ is the unique myopic best reply to beliefs concentrated in $D_a$ to show that with positive probability the agent will stick to action $a$ forever.

**Corollary 3.** Suppose that the agent has causation neglect, and that for all actions $a \in A$, supp $p^*_a = Y$. Then every uniformly strict Berk-Nash equilibrium is positively attracting.

Esponda, Pouzo, and Yamamoto (2019) shows that for every uniformly strict Berk-Nash equilibrium $a$, there exists at least a full support prior under which the policy converges to $a$.

<table>
<thead>
<tr>
<th>$a$</th>
<th>$a = 1$</th>
<th>$a = 2$</th>
<th>$a = 3$</th>
<th>$H(p^*_a, \cdot)$</th>
<th>$A^m(\cdot)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>1 2 3</td>
<td>1 2 3</td>
<td>1 2 3</td>
<td>$a = 1$ $a = 2$ $a = 3$</td>
<td></td>
</tr>
<tr>
<td>$u$</td>
<td>1 0 0</td>
<td>0 1 0</td>
<td>0 0 1</td>
<td>0.1 0.1 0.8</td>
<td>0.9 0.1 0.2</td>
</tr>
<tr>
<td>$p^*$</td>
<td>0.1 0.9 0</td>
<td>0.9 0.1 0</td>
<td>0.1 0.1 0.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p^1$</td>
<td>0.5 0.3 0.2</td>
<td>0.5 0.3 0.2</td>
<td>0.5 0.3 0.2</td>
<td>1.15 0.74 2.03</td>
<td></td>
</tr>
<tr>
<td>$p^2$</td>
<td>0.3 0.5 0.2</td>
<td>0.3 0.5 0.2</td>
<td>0.3 0.5 0.2</td>
<td>0.74 1.15 2.03</td>
<td></td>
</tr>
<tr>
<td>$p^3$</td>
<td>0.1 0.1 0.8</td>
<td>0.1 0.1 0.8</td>
<td>0.1 0.1 0.8</td>
<td>2.3 2.3 0.64</td>
<td></td>
</tr>
</tbody>
</table>

The table shows the subjective likelihoods assigned to the different outcomes for each action, along with the KL divergence from the empirical frequency.
with positive probability. Proposition 4 shows that under causation neglect, every uniformly strict Berk-Nash equilibrium has positive probability of being the limit behavior starting from every initial prior, even if the prior does not have full support. Frick, Iijima, and Ishii (2019b) focuses on the probability of convergence to a particular Berk-Nash equilibrium. They provide sufficient conditions for this probability to be arbitrarily high for beliefs sufficiently close to those that support the equilibrium. In contrast, Proposition 4 considers priors that can be far from the equilibrium belief.

4.2.2 Subjective Bandit Problems

Recall that in a subjective bandit problem, the agent’s prior \( \mu_0 \) is a product measure \( \mu_0 = \prod_{a \in A} \mu_a \). Each \( \mu_a \) is the belief about the distribution over outcomes induced by action \( a \).

Proposition 5. In a subjective bandit problem there exists a \( \hat{\beta} < 1 \) such that if \( \beta \geq \hat{\beta} \) then under any optimal policy \( \pi \), every uniformly strict Berk-Nash equilibrium is positively attracting.

The proof uses the fact that patient agents experiment with actions that they believe might give them a higher payoff. The conclusion of the proposition is false for myopic agents even in the correctly specified case, where the Berk-Nash equilibria correspond to the self-confirming equilibria, and with probability 1 the agent may always play whichever action is myopically optimal given their initial beliefs.

The requirement of a uniformly strict Berk-Nash equilibrium is very demanding in a perceived bandit setting, as the Kullback-Leibler divergence between the true and subjective outcome distributions induced by action \( b \) does not constrain the “off-path” beliefs about the consequences of other actions, and very optimistic off-path beliefs can make some other action a better reply.

However, if the equilibrium is weakly identified, we can relax uniformly strict to strict.

Definition 8. A Berk-Nash equilibrium \( b \) is weakly identified (Esponda and Pouzo, 2016) if for all \( p, p' \in D_b \) we have \( p_b = p'_b \).

Weak identification guarantees that once behavior stabilizes on action \( b \), there is no additional updating about the relative likelihood of the KL minimizing outcome distributions. Under causation neglect, weak identification is a relatively strong condition, as it requires that the KL minimizer is unique. Weak identification is significantly weaker in subjective
bandits, as it only requires the existence of a unique conceivable outcome distribution \( q \) that best matches \( p^*_b \), without imposing any restrictions on what the agent believes about the consequences of other actions.

**Proposition 6.** For every (perceived) bandit problem there is a \( \tilde{\beta} < 1 \) such that if the discount factor \( \beta \geq \tilde{\beta} \), then under any optimal policy \( \pi \), every weakly identified strict Berk-Nash equilibrium is positively attracting.

### 4.2.3 Strongly Supermodular environments

**Definition 9.** We say that the decision problem is *strongly supermodular* if we can strictly order the space of actions \((A, >)\), outcomes \((Y, >)\), and the set of beliefs over conceivable distributions \((\mathcal{M}, >)\) so that:

1. \( u \) is strictly supermodular in \( a \) and \( y \);
2. if \( p, p' \in \mathcal{M} \) and \( p > p' \), then for all \( a \in A \), and \( y \in Y \) we have
   \[ p_a (\{y' : y' > y\}) > p'_a (\{y' : y' > y\}). \]

**Proposition 7.** In a strongly supermodular decision problem, if \( p^*_a \) (resp. \( p^*_\tilde{a} \)) has full support, and the highest action \( \bar{a} \) (resp. the lowest action \( a \)) is a uniformly strict Berk-Nash equilibrium, then \( \bar{a} \) (resp. \( a \)) is positively attracting.

Strong supermodularity implies that for each action there is a set of outcome realizations, the highest \( y \)'s, that if observed supports the use of action \( \bar{a} \). A finite number of such realizations will induce the agent to use action \( \bar{a} \), and from that moment our Proposition 7 guarantees that there is positive probability that the agent will use \( a \) forever.

### 5 Signals

Suppose each period the agent observes a finite-dimensional vector \( s \in S \) before taking an action. Thus the analog of an action in the previous sections is now a *strategy*, namely a map \( \sigma : S \to A \) from signals to actions. Signals may be payoff relevant, so now utility is a map \( u : A \times Y \times S \to \mathbb{R} \), and may be useful for predicting the outcome distributions, i.e., now \( p_{a,s} \in \Delta(Y) \) depends both on this period action and signal.

We are interested in the case where the distribution \( p_{a,s} \) on outcomes depends on both \( a \) and \( s \), but the agent mistakenly thinks that some or all of the dimensions of \( s \) are irrelevant.
Here we expand the probability space of our basic model in the obvious way: The sample space $\Omega = S \times (Y^\infty)^A$ consists of infinite sequences of signal and action dependent outcome realizations $(s_k, x_{a,s',k})_{k \in \mathbb{N}, a \in A, s' \in S}$ and $x_{a,s',k}$ determines the outcome when the agent takes the action $a$ for the $k$-th time after $s$.\footnote{We specify all the details of the sample space and its probability measure in Appendix 8.3.}

To complete the model we also need to specify the distribution of signals. We focus here on the case where the distribution of $s$ is fixed (iid) with distribution $\zeta$, which is known to the agent, as in Esponda and Pouzo (2016).

**Subjective Beliefs of the Agent** The agent correctly believes that the map from actions and signals to probability distributions over outcomes is fixed, but they are uncertain about the distribution each signal and action pair induces. Let $P = (\Delta(Y))^{[A] \times [S]} \subset \mathbb{R}^{[Y] \times [A] \times [S]}$ be the space of all signal and action dependent outcome distributions. The agent’s uncertainty is captured by a prior belief $\mu_0 \in \Delta(P)$, with again $\Theta = \text{supp} \mu_0$. The only restriction is the generalization of Assumption 1.

**Assumption 2.** For all $p \in \Theta$ and $\varepsilon > 0$, there exists $p' \in \Theta$ with $||p' - p|| < \varepsilon$ such that for all $a \in A, s \in S$, if $p^*_{a,s}(y) > 0$ then $p'_{a,s}(y) > 0$.

Denote by $\mu_t(\cdot \mid (s^t, a^t, y^t)) \in \Delta(P)$ the agent’s subjective belief obtained using Bayes rule after observing the sequence of signals and outcomes $(s^t, y^t)$ when taking the actions $a^t$, so that

$$\mu_t(\Psi \mid (s^t, a^t, y^t)) = \frac{\int_{p \in \Psi} \prod_{\tau=1}^{T} p_{a^t,s^t} (y_{\tau}) d\mu_0(p)}{\int_{p \in P} \prod_{\tau=1}^{T} p_{a^t,s^t} (y_{\tau}) d\mu_0(p)}.$$ \hspace{1cm} (4)

The relevant set of “closest beliefs to the truth” is now

$$D_\sigma = \arg\min_{p \in \Theta} \sum_{s \in S} \zeta(s) H(p^*_{\sigma(s),s}, p_{\sigma(s),s}).$$ \hspace{1cm} (5)

We use this modified definition of $D$ to extend the definition of Berk-Nash equilibrium and uniformly strict Berk-Nash equilibrium to this more general setting.

**Definition 10** (Berk-Nash Equilibrium).

(i) Strategy $\sigma$ is a **Berk-Nash equilibrium** if there exists a belief $\nu \in \Delta(D_\sigma)$ such that $\sigma$ is myopically optimal given $\nu$.

(ii) Strategy $\sigma$ is a **uniformly strict Berk-Nash equilibrium** if $\sigma$ is the unique myopic best reply to any belief in $\nu \in \Delta(D_\sigma)$.
Proposition 8. Suppose the agent uses an optimal myopic policy. Then if the strategy prescribed by the policy converges to $\sigma$ with positive probability, then $\sigma$ is a Berk-Nash equilibrium.

The proof associates each action contingent outcome distribution $p$ with the joint distribution $p_\sigma \in \Delta(S \times Y)$ over signals and outcomes that is induced by $p$ and the use of strategy $\sigma$. We then use Berk’s Theorem for the belief process in this augmented environment to argue that the policy converges to $\sigma$, then beliefs concentrate around $D_\sigma$. We finally use the 0-1 Law to show that if $\sigma$ is not optimal with respect to some beliefs concentrated in $D_\sigma$, almost surely the optimal myopic policy will prescribe an action different from the one dictated by $\sigma$, a contradiction.

5.1 Signal Neglect

When the agent has “signal neglect,” they think the distribution of $y$ given $a$ is independent of $s$, so the only reason why the signals matter for the agent is that they may enter into the payoff function directly.\(^{18}\) This is very similar to the case where the agent estimates the same distribution for all actions, but here for each $a$ that is played infinitely often, Berk-Nash equilibrium requires that beliefs minimize the divergence with the signal-independent belief

$$D_\sigma = \arg\min_{\bar{p} \in \text{supp}(\mu)} H(\bar{p}_\sigma, p),$$

where $\bar{p}_\sigma = \sum_s \zeta(s)p_{\sigma(s),s}(y)$.

Example 6 illustrates the long-run biases that can be induced by signal neglect. There, we study the choice of a seller who receive a signal about the market attendance in the current period, and can decide whether to undertake an investment that may boost sales, with the outcome $y$ the fraction of market participants who buy. The seller does not realize that when more consumers show up, a lower fraction of them buy, and we show that this can lead to persistent underinvestment when the market attendance is high.

Definition 11 (Positively Attracting). The strategy $\sigma \in A^S$ is positively attracting if for every optimal policy $\pi$

$$P_\pi\left[\lim_{t \to \infty} \pi(s^{t+1}, a^{t}, y^t) = \sigma(s^{t+1})\right] > 0.$$  

\(^{18}\)If the agent thinks, correctly or not, that the signals are not payoff relevant, signal neglect reduces to replacing $p^*$ with the compound distribution over outcomes. With this change all the previous propositions continue to hold.
Proposition 9. Under causation and signal neglect, if sup $p^*_{a,s} = Y$ for all $(a,s) \in A \times S$, then any uniformly strict Berk-Nash equilibrium $\sigma$ is positively attracting.

The proof structure is very similar to Proposition 4 because with signal neglect we can apply our extension of the Diaconis and Freedman (1990) result to the uncontingent empirical distribution.

5.2 Markov Decision Problems

If the agent’s action influences the signal, then the true model is that of a Markov decision problem. Even if the agent ignores this, the evolution of their beliefs and actions becomes more complicated. And if the agent is aware of it, and tries to solve a Markov decision problem as in Esponda and Pouzo (2019) then the problem is yet more complex. We hope to have more to say about this in the next version of the paper...

6 Economic Examples

6.1 Learning in Large Population Games

Our framework can be used as the building block for steady-state learning in large population games. We now highlight how to map our framework in that setting and why the biases we consider are relevant there. Consider a finite game with a set of players roles $I$, and action space $A = \times_{i \in I} A_i$. There is a continuum of agents in each player role $i \in I$. Every period, the agents are matched uniformly at random to play the game. Agents discount the future at rate $\hat{\beta}$, and they have a probability $\gamma$ of surviving into next period.

In a steady state, the problem faced by an agent in population $i$ is formally equivalent to the one we have considered in the previous sections: the agent correctly believes they are facing a stationary environment, and they realize that they do not affect the next period distribution of opponents’ strategies. Here, the finite set of actions $A$ is $A_i$, $Y = A_{-i}$, $u$ is the payoff function of player $i$, and the actual discount factor of the agent is $\beta = \hat{\beta} \gamma$, and for simplicity, we assume that at the end of the period the agent perfectly observes the strategy of the opponents.

The objective probability distribution $p^* \in P = \Delta (Y)^A$ is determined by the equilibrium behavior of the opponents. The distribution over opponents actions may depend on the agent choice because the game may be dynamic, and the opponent may have the possibility to (partially) observe the behavior of the agent before moving. With this interpretation...
in mind, we can map our particular misspecification into wrong beliefs about the game structure. Causation neglect corresponds to the bias of an agent who thinks they are playing a simultaneous game, whereas in reality the opponents observe their choice before moving. Subjective bandit problems are situations in which the agent realizes that the opponents observe their action before moving, and the agent has an independent belief about the responses to different strategy. Finally, the extension to signals allow to partially deal with games of incomplete information. There signal neglect corresponds to a situation in which the agent observes a signal and they think that the game they are playing has private value: the observed signal may be payoff relevant, but does not convey information about what the other players are going to do.

In all of these situations, our results help characterize the possible limit actions of an agent who faces a given steady-state play by the opponents described by \( p^* \). This is a building block for the analysis of which steady-states can arise in the limit where agents become infinitely patient and long-lived. That said, extensive-form games may not have strict equilibria, so many of our Propositions will not apply to them. However, it may be possible to extend some conclusions to equilibria that are on-path strict in the sense of Fudenberg and He (2020). Also, games need not have pure-strategy equilibria, but it may be possible to apply our methods to setting where each agent plays deterministically and different agents in the same player role chose different actions.\(^{19}\)

### 6.2 Other Examples

**Example 4 (Central Bank Policy).** A central bank decides between three actions:

- **f**: keep a flexible exchange rate with the dollar,
- **c**: peg the currency to the dollar imposing some controls to the capital, or
- **g**: peg the currency without imposing controls.

The outcome has three binary components, \( y = (y^e, y^s, y^{cf}) \), where \( y^e \) says whether the economy is in expansion in the current period, \( y^s \) whether there is a speculative attack on the currency, and \( y^{cf} \) if there is capital flight. The central bank believes that the three components are distributed according to independent binomial distributions that do not depend on the action, with

\[
p(y) = \eta^e y^e \eta^s y^s \eta^{cf} y^{cf} (1 - \eta^e)^{1-y^e} (1 - \eta^s)^{1-y^s} (1 - \eta^{cf})^{1-y^{cf}}.
\]

\(^{19}\)Alternatively we could enrich the model with payoff perturbations as in Fudenberg and Kreps (1993) and Esporda and Pouzo (2016), but this makes the effective action space infinite so our current results would not directly apply.
The payoff function is given by

\[
\begin{align*}
    u(f,y) &= y^e \\
    u(c,y) &= \frac{3}{2} y^e - y^s \\
    u(g,y) &= 2y^e - \frac{y^{cf}}{2} - \frac{y^s}{2}.
\end{align*}
\]

This specification of the payoffs captures the fact that a fixed exchange rate facilitates trade and foreign investment when the currency isn’t attacked.\(^{20}\) In this setting, Proposition 4 guarantees that if \(p^*\) has full support, every uniformly strict Berk-Nash equilibrium has positive probability of arising as the limit outcome.

---

Here pegging the currency is a uniformly strict Berk-Nash equilibrium but it is not positively attractive.\(^{21}\)

**Example 5** (Role of Assumption 1). Suppose there are two actions \(a\) and \(b\), and two outcomes \(Y = \{0,1\}\), and let \(u(a,0) = u(b,1) = 1 - u(a,1) = 1 - u(b,0)\). We identify the elements of \(\Delta(Y)\) with the probability they assign to outcome 1. Let \(p^*_a = \frac{2}{3}\) and \(p^*_b = 1\). Suppose that the agent believes that the outcome distribution does not depend on the action, so they have correlation neglect, and let \(\Theta = \{\frac{1}{3}, 1\}\). Here \(b\) is the unique Berk-Nash equilibrium, and it is uniformly strict. However, if the prior assigns a sufficiently high probability to 1/3,

\(^{20}\)See e.g., Frankel, Schmukler, and Servén (2004).

\(^{21}\)When the currency is pegged to the dollar, \(p^2\) is indistinguishable from \(\tilde{p}^1\), and \(p^2\) is indistinguishable from \(p^1\), so the bank doesn’t update their beliefs. Thus there is probability 0 of converging to \(g\) when the prior assigns sufficiently high probability to \(\tilde{p}^2\) and \(p^2\).
the agent will start playing $a$, and with positive probability they will observe outcome 0 in the first period. But after this observation, the posterior assigns probability one to $p = 1/3$ and the actions will converge to $a$.

**Example 6** (Signal Neglect). A seller in a physical marketplace can hire one shop assistant to work for the day $a_H$ or not hire anyone $a_N$. The outcome $y \in Y$ is the percentage of consumers in the marketplace that buy the good, with two possibilities, $y_h > y_l$.

Before choosing whether to hire, the agent observes the the number of people at the market that day $s \in \{s_h, s_l\}$, with $s_h > s_l$. The payoff function is $u(a, y, s) = sy - 1_{c=a_H}$. The seller realizes that the signal is payoff relevant, but falsely believes that it does not provide any information about the outcome. The agent is uncertain about how useful it is to hire a shop assistant, and in particular they do not know whether hiring is ineffective, i.e., for all $a \in A, y \in Y$, $p_a(y) = 1/2$, or if it is not, i.e., $p_{a_H}^*(y_H) = 3/4$ and $p_{a_N}^*(y_H) = 1/4$.

The fraction of consumers who buy varies with the signal: On days with fewer consumers, the ones that actually come to the market are more likely to purchase the good. Formally:

$$p_{s_H,a_H}^*(y_H) = 1/2, \quad p_{s_H,a_N}^*(y_H) = 1/4, \quad p_{s_L,a_H}^*(y_H) = 3/4, \quad p_{s_L,a_N}^*(y_H) = 1/2.$$  

Let $\frac{s_l(y_h - y_l)}{4} < 1 < \frac{s_h(y_h - y_l)}{4}$, so that it is not optimal to hire a shop assistant after $s_L$, and it is optimal to hire an assistant after $s_H$. Then the only Berk-Nash equilibrium is that the shop assistant is never hired.\(^{22}\)

7 Conclusion

In many economically relevant settings it seems plausible that agents misunderstand some aspects of the world. For this reason it is important to understand what beliefs such agents will develop and how they will behave. This paper provides sharp characterizations of what actions arise as the long-run outcomes of misspecified learning. We show that all uniformly strict Berk Nash equilibria are stable, and that under a mild condition only uniform Berk Nash equilibria can be stable. Moreover we show that when the support of the agent’s prior

\(^{22}\)If the agent followed the optimal strategy, they would observe the same frequency of sales in days with $s = s_H$ and with the shop assistant hired as in days with $s = s_L$ and without the shop assistant: $p_{s_H,a_H}^*(y_H) = 1/2 = p_{s_L,a_N}^*(y_H)$ This holds because the shop assistant offsets the lower per-customer demand on days with high attendance. However, the agent think this confirms that the signal is irrelevant for the outcome distribution. end up believing that the shop assistant is useless, i.e. $\mu_t(p) \to 1$. Since the myopic best reply to $\delta_p$ is to never hire the shop assistant, by Proposition 8 this suboptimal action is the only possible limit action.
is finite, play can only converge to uniform Berk-Nash equilibria.\footnote{Note that the uniformity issue that we address cannot arise in a correctly specified model, where the agent always learn the outcome distribution induced by the equilibrium action. Note also that Of our results do not imply that actions converge.} Our work thus suggests uniformity should be imposed as a refinement of Berk-Nash equilibrium. We then provide the first sufficient conditions for an action to be positively attracting under misspecified learning. Here we highlight the role played by the correlation that the agent perceives between the outcome distributions associated with different actions.

In future work we hope to extend our analysis to Markov decision problems, as in Esponda and Pouzo (2019) and to misspecified learning in multiplayer games, as in Eyster and Rabin (2005), Jehiel (2005), and Jehiel and Koessler (2008).

8 Appendix

The structure of the appendix is as follows. Section 8.1 establishes some preliminary lemmas. Section 8.2 contains the results of the main text for the models that do not have signals, except for the first part of Proposition 1 which will follow immediately from the more general Proposition 8. Finally, Section 8.3 contains some technical details about the probability space and the proofs of the results for the model with signals.

8.1 Preliminary lemmas and definitions

For every \( p \in P \) and every policy \( \pi \) denote as \( \mathbb{E}_{p,\pi} \) the probability distribution over action and outcome sequences that is induced by policy \( \pi \) under outcome distribution \( p \). For convenience, for every policy \( \pi \) and initial belief \( \nu \) we will work with the agent’s normalized value throughout, which is

\[
V(\pi, \nu) = \frac{\int_P \mathbb{E}_{p,\pi} \left[ \sum_{t=1}^{\infty} \left[ \beta^{t-1} u(a_t, y_t) \right] \right] d\nu(p)}{1 - \beta}.
\]

We first bound the difference between the value of using action \( a \) and the value of any other action in terms of their expected utility given that beliefs are concentrated around the outcome distributions \( D_a \) that minimize the Kullback-Leibler divergence from the correct distribution \( p^*_a \) induced by \( a \). To do this we will make use of the inverse projection operator \( J_a : \Delta(Y) \to 2^P \) which maps a distribution over outcomes \( q \in \Delta(Y) \) to the set of all conceivable action contingent distributions with the same marginal distribution associated
with action $a$

$$J_a(q) = \{ p \in \Theta : p_a = q \}.$$ 

Define the set $\hat{D}_{\varepsilon,a}$ as all outcome distributions such that the marginal distribution with respect to action $a$ is at most $\varepsilon$ away from a KL minimizer

$$\hat{D}_{\varepsilon,a} = \{ p \in \Theta : \text{there exists } q \in D_a \text{ with } |q_a - p_a| \leq \varepsilon \},$$

and the set of beliefs over conceivable distributions that assign at least probability $1 - \varepsilon$ to $\hat{D}_{\varepsilon,a}$

$$\mathcal{M}_{\varepsilon,a} = \{ \nu \in \mathcal{M} : \nu(\hat{D}_{\varepsilon,a}) \geq 1 - \varepsilon \}.$$ 

**Lemma 1.** If $a \in A$ is a uniformly strict Berk-Nash equilibrium, for every optimal policy $\pi$, there exists an $\hat{\varepsilon} > 0$ such that for all $\varepsilon < \hat{\varepsilon}$

$$\nu \in \mathcal{M}_{\varepsilon,a} \implies \pi(\nu) = a.$$

**Proof.** Let $\pi^a$ denote the policy that prescribes to always play $a$. Define $G(\varepsilon)$ as the minimal gain from playing $a$ forever instead of using (one of) the best policy $\tilde{\pi}$ that does not play $a$ at a prior belief $\nu$ in $\mathcal{M}_{\varepsilon,a}$

$$G(\varepsilon) = \min_{\tilde{\pi} : \tilde{\pi}(\nu) \neq a} \min_{\nu \in \mathcal{M}_{\varepsilon,a}} (V(\pi^a, \nu) - V(\tilde{\pi}, \nu))$$

$$\geq (1 - \beta) \min_{a' \neq a} \min_{\nu \in \mathcal{M}_{\varepsilon,a}} \left( \int_P \mathbb{E}_{p_a} [u(a, y)] d\nu(p) - \int_P \mathbb{E}_{p_{a'}} [u(a', y)] d\nu(p) \right)$$

$$+ \beta \min_{a' \in A} \min_{\nu \in \mathcal{M}_{\varepsilon,a}} \left( \int_P \mathbb{E}_{p_a} [u(a, y)] d\nu(p) - \int_P \mathbb{E}_{p_{a'}} [u(a', y)] d\nu(p) \right),$$

where the inequality follows from Jensen inequality and the concavity of the minimum. By the Maximum Theorem, the last expression is continuous in $\varepsilon$, and since $a$ is a uniformly strict Berk-Nash equilibrium, $G(0) > 0$, there is an $\hat{\varepsilon}$ such that if $\varepsilon \leq \hat{\varepsilon}$, $G(\varepsilon) > 0$. This implies that $\tilde{\pi}$ is sub-optimal, so any optimal policy $\pi$ it must be that $\nu \in \mathcal{M}_{\varepsilon,a}$ implies that $\pi(\nu) = a$, which proves the lemma.

The next lemma consider a compact set of outcome distributions $\Psi \subseteq \Delta(Y)$ and an outcome distribution $p' \in \Delta(Y)$. It shows that if every distribution in $\Psi$ dominates $p'$ according to the Renyi divergence with parameter $l$, then the $l$-th power of the odds ratio of $\Psi$ to a neighborhood of $p'$ is a supermartingale. This extends an argument of Fudenberg...
and Levine (1992) to take into account misspecification and to allow for \( l < 1 \).

**Lemma 2.** Let \( \Psi \) be a compact subset of \( \Delta(Y) \), \( \rho' \in \Delta(Y) \), \( l \in (0,1) \) and \( \rho^* \in \Delta(Y) \) such that

\[
\forall \rho \in \Psi \quad \sum_{y \in Y} \rho^*(y) \left( \frac{\rho(y)}{\rho'(y)} \right)^l \geq 1.
\]

Then, there is \( \varepsilon_{\rho'} > 0 \) such that for all \( \nu \in \Delta(\Delta(Y)) \), if we let

\[
\nu(C \mid y) = \frac{\int_{y \in C} q(y) d\nu(q)}{\int_{y \in \Delta(Y)} q(y) d\nu(q)},
\]

then

\[
\sum_{y \in Y} \rho^*(y) \left[ \left( \frac{\nu(\Psi \mid y)}{\nu(B_{\varepsilon_{\rho'}}(\rho') \mid y)} \right)^l \right] \geq \left( \frac{\nu(\Psi)}{\nu(B_{\varepsilon_{\rho'}}(\rho') \mid y)} \right)^l.
\]

**Proof.** Without loss of generality, we can assume that \( \nu(\Psi) > 0 \). By the Maximum Theorem and equation (6), there is \( \varepsilon_{\rho'} > 0 \) such that for all \( \tilde{\nu} \in \Delta\left( B_{\varepsilon_{\rho'}}(\rho') \right) \),

\[
\sum_{y \in Y} \rho^*(y) \left( \frac{p(y)}{\int_{\Delta(B_{\varepsilon_{\rho'}}(\rho'))} q(y) d\tilde{\nu}(q)} \right)^l \geq 1
\]
Then,

$$\sum_{y \in Y} p^*(y) \left[ \frac{\nu(\Psi | y)}{\nu(B_{\epsilon', \nu} (p') | y)} \right]^l$$

$$= \sum_{y \in Y} p^*(y) \left( \frac{\int_{\Delta (B_{\epsilon', \nu} (p'))} \nu (\Psi) p' (y) d\frac{\nu(p')}{\nu(\Psi)} \int_{\Delta (B_{\epsilon', \nu} (p'))} \nu (\Delta (B_{\epsilon', \nu} (p'))) q(y) d\frac{\nu(q)}{\nu(\Delta (B_{\epsilon', \nu} (p')))} }{\int_{\Delta (B_{\epsilon', \nu} (p'))} \nu (\Delta (B_{\epsilon', \nu} (p'))) q(y) d\frac{\nu(q)}{\nu(\Delta (B_{\epsilon', \nu} (p')))} } \right)^l$$

$$\geq \sum_{y \in Y} p^*(y) \int_{\Psi} \left( \frac{\nu (\Psi) p' (y) \int_{\Delta (B_{\epsilon', \nu} (p'))} \nu (\Delta (B_{\epsilon', \nu} (p'))) q(y) d\frac{\nu(q)}{\nu(\Delta (B_{\epsilon', \nu} (p')))} }{\int_{\Delta (B_{\epsilon', \nu} (p'))} \nu (\Delta (B_{\epsilon', \nu} (p'))) q(y) d\frac{\nu(q)}{\nu(\Delta (B_{\epsilon', \nu} (p')))} } \right)^l d\frac{\nu(p')}{\nu(\Psi)}$$

$$= \int_{\Psi} \frac{\nu (\Psi)^l}{\nu (\Delta (B_{\epsilon', \nu} (p')))} \sum_{y \in Y} p^*(y) \left( \frac{\nu(q)}{\nu(\Delta (B_{\epsilon', \nu} (p'))) } \right) \sum_{y \in Y} p^*(y) \left( \frac{\nu(q)}{\nu(\Delta (B_{\epsilon', \nu} (p'))) } \right) \int_{\Psi} \frac{\nu(p')}{\nu(\Psi)}$$

$$\geq \int_{\Psi} \frac{\nu (\Psi)^l}{\nu (\Delta (B_{\epsilon', \nu} (p')))} \sum_{y \in Y} \int_{\Psi} \frac{\nu(p')}{\nu(\Psi)}$$

$$\sum_{y \in Y} p^*(y) \left[ \frac{1 - \nu(\Psi' | y)}{\nu(\Psi | y)} \right]^l \leq \left( \sum_{y \in Y} p^*(y) \left[ \frac{1 - \nu(\Psi' | y)}{\nu(\Psi | y)} \right] \right)^l \leq \left( \frac{1 - \nu(\Psi)}{\nu(\Psi)} \right)^l.$$
Proof. As noted by Frick, Iijima, and Ishii (2019b) in their Lemma 3, a simple continuity argument shows that for each KL minimizer \( q \in D_a \) and every outcome distribution \( p' \notin \hat{D}_{\varepsilon,a} \), there exists an \( l(q,p') \) such that \( f_l(q,p') > 1 \) for all \( l \leq l(q,p') \). We will now prove that there exists a uniform \( l \) that works for every \( q \in D_a \) and \( p' \in \hat{D}_{\varepsilon,a} \).

Suppose by way of contradiction that there was no \( \frac{l}{2} \in (0,1) \) such that \( f_l(q,p') > 1 \) for all \( q \in D_a \) and \( p' \notin \hat{D}_{\varepsilon,a} \). Then, we can define a sequence \((q_n,p'_n)\) such that \( f_{\frac{l}{2}}(q_n,p'_n) \leq 1 \), and sequential compactness of \( D_a \times \{ p \in \mathcal{M} : p \notin \hat{D}_{\varepsilon,a} \} \) guarantees that this sequence has an accumulation point \((q,p')\). However, for, \( n > \frac{1}{l(p,p')}, f_{\frac{l}{2}}(q_n,p'_n) \leq 1 \) implies \( f_{l(q,p')} \)\( (q_n,p'_n) \leq 1 \), but then continuity of \( f_{l(q,p')} \) leads to a contradiction with \( f_{l(q,p')} > 1 \).

The following lemma is about the concentration of beliefs. The lemma considers the beliefs about the outcome distribution corresponding to a given action, i.e. to elements of \( \Delta(Y) \), as opposed to elements of \( \times_{a \in A} \Delta(Y) \subseteq \mathbb{R}^{|Y| \times |A|} \), so we will lighten notation by working in this smaller space.

Let \( \nu \in \Delta(\Delta(Y)) \) be a belief over probability distributions on \( Y \), and let

\[
M_{\varepsilon,\nu}(\bar{q}) = \left\{ q' \in \Delta(Y) : \exists q'' \in \text{argmin}_{q \in \text{supp} \nu} H(\bar{q}, q), ||q' - q''||_{\infty} < \varepsilon \right\}
\]

be the distributions in \( \text{supp} \nu \) that are within \( \varepsilon \) of a distribution \( q'' \) that minimizes the Kullback-Leibler divergence with the given \( \bar{q} \).\(^{24}\) We will show that beliefs about the outcome distribution induced by any action \( a \) concentrate around \( M_{\varepsilon,\text{supp} \mu_a} \), the distributions that best fit the empirical frequency of outcomes generated by \( a \). Importantly, this result does not require that either actions or empirical frequencies converge. It is based on arguments made in Diaconis and Freedman (1990), who considered agents with full support beliefs. It will be important in what follows that these results apply pathwise.

**Lemma 4.** Let \( \nu_0 \in \Delta(\Delta(Y)) \) and suppose that for every \( t \in \mathbb{N} \), \( \Psi \subseteq \Delta(Y) \), and sequence of outcomes \( y^t \in Y^t \)

\[
\nu_t(\Psi | y^t) = \frac{\int_{\Psi} \prod_{t=1}^t q(y_r) d\nu_0(q)}{\int_{\Psi} \prod_{t=1}^t q(y_r) d\nu_0(q)}.
\]

Then there is a function \( g : \Delta(Y) \times [0,1] \to \mathbb{R}_+ \) that is upper-semicontinuous in the second

\(^{24}\)Note that the \( \text{argmin} \) in this definition need not be continuous because \( \text{supp} \nu \) is not required to be convex.
argument and such that for every \( \varepsilon > 0 \)

\[
p'(y) = \frac{\sum_{t=1}^{1} y_{t} = y}{t} \implies \frac{\nu_{t} (M_{\varepsilon \nu_{0}} (p') | y')} {1 - \nu_{t} (M_{\varepsilon \nu_{0}} (p') | y')} \geq K_{\varepsilon,p'} e^{g(p',\varepsilon)},
\]

where \( K_{\varepsilon,p'} > 0 \).

**Proof.** Let \( p'(y) = \frac{\sum_{t=1}^{1} y_{t} = y}{t} \) and fix \( \varepsilon > 0 \). To ease notation, in this proof for every \( \varepsilon > 0 \), we let \( M(\varepsilon) = M_{\varepsilon \nu_{0}} (p') \). Define

\[
g(p',\varepsilon) = \frac{\min_{p \notin M(\varepsilon)} H(p',p) - \min_{p \in \text{supp} \nu_{0}} H(p',p)}{2} > 0.
\]

Since \( g \) is upper-semicontinuous in the second argument (see, e.g., Lemma 17.30 in Aliprantis and Border, 2013), for every \( \bar{p} \in \arg \min_{p \in \text{supp} \nu_{0}} H(p',p) \) there exists an \( \varepsilon_{\bar{p}} \leq \varepsilon \) such that for every \( \bar{q} \notin M(\varepsilon) \), implies

\[
\min_{p \notin M(\varepsilon)} H(p',p) - H(p',p) \geq g(p',\varepsilon).
\]

By the Maximum Theorem, \( \arg \min_{p \in \text{supp} \nu_{0}} H(p',p) \) is compact, and it is covered by the collection

\[
\bigcup_{\bar{p} \in \arg \min_{p \in \text{supp} \nu_{0}} H(p',p)} B^{\text{int}}(\bar{p}),(\varepsilon)
\]

and therefore it admits a finite subcollection. Let \( \varepsilon' \) be the minimum of the \( \varepsilon_{\bar{p}} \) indexing the subcollection. From the definition of \( \nu_{t} \), if we define \( K_{\varepsilon,p'} = \nu_{0} (M(\varepsilon'))^{2} \) we have that for all \( y' \) such that the corresponding empirical distribution is \( p' \),

\[
\frac{\nu_{t} (M(\varepsilon) | y')} {1 - \nu_{t} (M(\varepsilon) | y')} = \frac{\int_{M(\varepsilon)} \sum_{y \in Y} q(y)^{tp(y')} (1 - q(y))^{t(1-p(y'))} \, d\nu_{0}(q)} {\int_{\text{supp} \nu_{0} \setminus M(\varepsilon)} \sum_{y \in Y} q(y)^{tp(y')} (1 - q(y))^{t(1-p(y'))} \, d\nu_{0}(q)} \geq \frac{\int_{M(\varepsilon')} \exp(-tH(p',q)) \, d\nu_{0}(q)} {\exp(-t \min_{p \notin M(\varepsilon')} H(p',p))}
\]

\[
= \nu_{0} (M(\varepsilon')) \int_{M(\varepsilon')} \exp(t \min_{p \notin M(\varepsilon')} H(p',p) - tH(p',q)) \, d\nu_{0}(q)
\]

\[
\geq \nu_{0} (M(\varepsilon'))^{2} e^{g(p',\varepsilon)}
\]

\[
= K_{\varepsilon,p'} e^{g(p',\varepsilon)}.
\]
The lemma shows that whether or not actions converge, the agent becomes confident that the true outcome generating process lies in a neighborhood of the observed empirical frequency, with a bound that is uniform over the realizations of those frequencies.

**Remark 2.** Appendix 8.4 shows that a uniform version of Lemma 4 holds in the case of binary outcomes (i.e. $y \in Y = \{0, 1\}$) and sufficiently small misspecification.

### 8.2 Proof of results stated in the text

**Proof of Proposition 1.** Proposition 1 (i) follows immediately from Proposition 8. Proposition 1 (ii) assumes that $|\Theta| = |\text{supp}\mu_0| < \infty$. Consider a Berk-Nash equilibrium $a$ that is not uniform. As $\Theta$ is finite $\bar{P}_a$ is finite; let $q^1, \ldots, q^{\bar{P}_a}$ denote the points in $\bar{P}_a$. Because $a$ is not a uniform Berk-Nash equilibrium, there is an outcome distribution $q^j \in P_a$ such that for every $\nu \in \Delta (\{p \in \Theta : p_a = q\})$, the optimal action corresponding to that $\nu$ is not $a$, i.e. $a \notin A^m(\nu)$.

The proof has three steps. First, we construct a multidimensional process where entry $j$ is the relative log-likelihood of $q_j$ with respect to $q^1$, and show that is a random walk with increments that have positive definite covariance. Second, we use the Central Limit Theorem to show that the process converges (in distribution) to a normal distribution with strictly positive density. Finally, we combine the Kochen-Stone lemma with the Hewitt-Savage 0-1 Law to show that almost surely the posterior likelihood of $q^1$ becomes arbitrarily high.

Without loss assume that $j = 1$. Define for every $q^k \in \bar{P}_a$ and $k = 1, \ldots, |\bar{P}_a| - 1$

$$Z^k_t = \log \frac{\mu_t(\{p \in \Theta : p_a = q^{k+1}\})}{\mu_t(\{p \in \Theta : p_a = q^1\})}.$$  

Define $L^k_t = \log \frac{q^{k+1}(y_t)}{q^1(y_t)}$. We have that $Z_t = Z_0 + \sum_{s=1}^t L_t$ and that

$$\mathbb{E}[L^k_t \mid (Z_s)_{s \leq t}] = \mathbb{E}[Z^k_t - Z^k_{t-1}] = \sum_y p^*_a(y) [\log(q^{k+1}(y)) - \log(q^1(y))] = 0, \quad (7)$$

where the last equality is due to the fact that $q^1$ and $q^{k+1}$ are both KL minimizers, i.e. $q^1, q^{k+1} \in \bar{P}_a$. The process $Z$ is a multi-dimensional random walk in $\mathbb{R}^{|\bar{P}_a|-1}$ and a by (7) a martingale. Furthermore, because $q^1$ is absolutely continuous with respect to $p^*_a$, the
increments $L_t$ have covariance matrix $\Sigma$ given by

$$
\Sigma_{ij} = \text{cov}(L^i, L^j) = \mathbb{E}[L^i L^j] = \sum_{y} p_a^*(y) \log \left( \frac{q_i^{i+1}(y)}{q_i^{i}(y)} \right) \log \left( \frac{q_j^{j+1}(y)}{q_j^{j}(y)} \right).
$$

We have that for any vector $v \in \mathbb{R}_{+}^{P_a}$ with $|v|_1 = 1$

$$
v^T \Sigma v = \sum_{i,j} v_i \Sigma_{ij} v_j = \sum_{i,j} v_i v_j \sum_{y} p_a^*(y) \log \left( \frac{q_i^{i+1}(y)}{q_i^{i}(y)} \right) \log \left( \frac{q_j^{j+1}(y)}{q_j^{j}(y)} \right)
\quad = \sum_{y} p_a^*(y) \sum_{i,j} v_i \log \left( \frac{q_i^{i+1}(y)}{q_i^{i}(y)} \right) v_j \log \left( \frac{q_j^{j+1}(y)}{q_j^{j}(y)} \right)
\quad = \sum_{y} p_a^*(y) \left( \sum_{i} v_i \log \left( \frac{q_i^{i+1}(y)}{q_i^{i}(y)} \right) \right)^2 \geq 0.
$$

The above expression equals zero if and only if for all $y \in Y$ we have that

$$
0 = \sum_{i=1}^{\lceil \tilde{P}_a \rceil - 1} v_i \log \left( \frac{q_i^{i+1}(y)}{q_i^{i}(y)} \right) \Rightarrow \log q_i^{i}(y) = \sum_{i=2}^{\lceil \tilde{P}_a \rceil} v_i \log(q_i^i(y)).
$$

By Jensen’s inequality we get that

$$
\log q_i^{i}(y) \leq \log \sum_{i=2}^{\lceil \tilde{P}_a \rceil} v_i q_i^i(y) \Rightarrow q_i^{i}(y) \leq \sum_{i=2}^{\lceil \tilde{P}_a \rceil} v_i q_i^i(y).
$$

As $\sum y q_i^i(y) = 1$ for all $i$ this implies

$$
q_i^{i}(y) = \sum_{i=2}^{\lceil \tilde{P}_a \rceil} v_i q_i^i(y).
$$

As $H$ is strictly convex in its second argument so this contradicts that $q_i^{i} \in \tilde{P}_a$. Thus, $v^T \Sigma v > 0$ for all $v \in \mathbb{R}_{+}^{\lceil \tilde{P}_a \rceil - 1}$ with $|v|_1 = 1$ and hence $\Sigma$ is positive definite.

Because $Z$ is a random walk with increments with positive definite covariance matrix and a martingale we can apply the central limit theorem and get that $\frac{Z_t}{\sqrt{t}}$ converges in distribution to a $\lceil \tilde{P}_a \rceil - 1$ dimensional Normal distribution with mean $\tilde{0}$ and covariance matrix $\Sigma$.

Fix an arbitrary number $c < 0$. We are interested in the events $E_t$ that $Z_t$ is coordinate-
wise less than $c^{25}$

$$E_t = \{ Z : Z_t \leq c \}.$$  

As $\frac{Z_t}{\sqrt{t}}$ converges to Normal random variable we have that

$$\lim_{t \to \infty} \mathbb{P}[E_t] = \lim_{t \to \infty} \mathbb{P}\left[ \frac{Z_t}{\sqrt{t}} \leq \frac{c}{\sqrt{t}} \right] = \mathbb{P}\left[ \tilde{Z} \leq 0 \right],$$

where $\tilde{Z}$ is a random variable that is Normally distributed with mean $\bar{0}$ and covariance matrix $\Sigma$. As $\Sigma$ is positive definite we have that this distribution admits a strictly positive density and hence $\mathbb{P}[\tilde{Z} \leq 0] > 0$. Consequently, it follows that

$$\sum_{t=1}^{\infty} \mathbb{P}[E_t] = \infty.$$  

We furthermore have that

$$\liminf_{t \to \infty} \frac{1}{t} \sum_{s=1}^{t} \sum_{r=1}^{t} \mathbb{P}[E_s \text{ and } E_r] = \liminf_{t \to \infty} \frac{1}{\frac{1}{t} \sum_{s=1}^{t} \sum_{r=1}^{t} \mathbb{P}[E_s]} \left( \frac{1}{\frac{1}{t} \sum_{s=1}^{t} \mathbb{P}[E_s]} \right)^2$$

$$\leq \liminf_{t \to \infty} \frac{1}{t} \sum_{r=1}^{t} \mathbb{P}[E_r] \left( \frac{1}{t} \sum_{s=1}^{t} \mathbb{P}[E_s] \right)^2$$

$$= \liminf_{t \to \infty} \frac{1}{\left( \frac{1}{t} \sum_{s=1}^{t} \mathbb{P}[E_s] \right)^2} = \frac{1}{\lim_{t \to \infty} \mathbb{P}[E_t]}$$

$$= \frac{1}{\mathbb{P}[\tilde{Z} \leq 0]}.$$  

It thus follows from the Kochen-Stone lemma (see Kochen, Stone, et al. (1964) or Exercise 2.3.20 in Durrett (2008)) that

$$\mathbb{P}\left[ \bigcap_{t=1}^{\infty} \bigcup_{s=t}^{\infty} E_s \right] \geq \mathbb{P}[\tilde{Z} \leq 0] > 0.$$  

As the above event is invariant under finite permutations of $L_t$ with different time indices Hewitt–Savage zero–one law (see, e.g., Theorem 8.4.6 in Dudley (2018)) implies that the probability of the event $\bigcap_{t=1}^{\infty} \bigcup_{s=t}^{\infty} E_s$ must equal zero or one. As the probability is strictly

---

25We use $\leq$ to denote component-wise comparison, i.e $Z_t \leq c$ means that $Z_i \leq c$ for all $i \in \{1, \ldots, |\bar{P}_s|-1\}$. 

---

34
positive it must equal one
$$\mathbb{P}\left[ \bigcap_{t=1}^{\infty} \bigcup_{s=t}^{\infty} E_s \right] = 1.$$  
This implies that the random walk \( Z \) falls below \( c \) with probability 1 infinitely often.

Denote now by \( q^1, \ldots, q^k \) the conceivable outcome distributions contingent to action \( a \), i.e., the elements of \( \{p_a \in \Delta(Y) : p \in \Theta\} \) and assume without loss that \( q^1, \ldots, q^{\bar{P}_a} \) are the KL minimizer. We consistently define \( Z^k_t = \log \frac{\mu_t(\{p \in \Theta : p_a = q^{k+1}\})}{\mu_t(\{p \in \Theta : p_a = q^k\})} \) and \( L^k_t = \log \frac{q^{k+1}(y_t)}{q^k(y_t)} \) for \( k \in \{|\bar{P}_a|, \ldots, |\Theta| - 1\} \). As

$$\mathbb{E}[L^k_t](Z_s)_{s \leq t} = H(p^*_a, q^1) - H(p^*_a, q^{k+1}) < 0$$

it follows that \( \lim_{t \to \infty} e^{Z^k_t} = 0 \) a.s. for all \( k \geq |\bar{P}_a| \). We have that

\[
\liminf_{t \to \infty} \log \frac{1 - \mu_t(\{p_a = q^k\})}{\mu_t(\{p_a = q^{k+1}\})} = \liminf_{t \to \infty} \log \frac{\sum_{k=1}^{\Theta-1} \mu_t(\{p \in \Theta : p_a = q^k\})}{\mu_t(\{p \in \Theta : p_a = q^{k+1}\})}
\]
\[
= \liminf_{t \to \infty} \log \sum_{k=1}^{\bar{P}_a-1} e^{Z^k_t} + \log \sum_{k=1}^{\bar{P}_a-1} e^{Z^k_t}
\]
\[
\leq \liminf_{t \to \infty} \sum_{k=1}^{\bar{P}_a-1} e^{\max_{k<|\bar{P}_a|} Z^k_t}
\]
\[
= \liminf_{t \to \infty} \log(\bar{P}_a - 1) + \max_{k<|\bar{P}_a|} Z^k_t.
\]

But, as we have argued before \( (Z^k_t)_{k <|\bar{P}_a|} \) falls below any threshold \( c \) infinitely often which implies that

$$\mathbb{P}\left[ \liminf_{t \to \infty} \max_{k<|\bar{P}_a|} Z^k_t = -\infty \right] = 1.$$  

Thus, the probability the agent assigns to \( q^1 \) as the outcome distribution contingent to action \( a \) gets arbitrary large

$$\mathbb{P}\left[ \limsup_{t \to \infty} \log \frac{\mu_t(\{p \in \Theta : p_a = q^1\})}{1 - \mu_t(\{p \in \Theta : p_a = q^1\})} = \infty \right] = \mathbb{P}\left[ \limsup_{t \to \infty} \mu_t(\{p \in \Theta : p_a = q^1\}) = 1 \right] = 1.$$  

As the action \( a \) is suboptimal when the agent assigns probability 1 to the outcome distributions in \( \{p \in \Theta : p_a = q^1\} \) it follows from continuity that there exists a threshold \( \mu^1 < 1 \) such that whenever the agent assigns probability greater than \( \mu^1 \) the action \( a \) is suboptimal.

As \( \limsup_{t \to \infty} \mu_t(\{p \in \Theta : p_a = q^1\}) = 1 \) it follows that the agent will always take an action
different from $a$ eventually
\[ P[a_t \neq a \text{ for some } t] = 1. \]

Thus, the action can not converge to $a$ with positive probability if $a$ is a Berk-Nash equilibrium, but not a uniform Berk-Nash equilibrium.

**Example 7.** To see that Proposition 1 (ii) does not hold for priors with an infinite support, modify example 1 as follow: Keep the same action space \{a, b\}, outcome space $Y = \{0, 1\}$, and now suppose the agent correctly believes that the action has no impact on the outcome distribution, and that $p^* = \frac{1}{2}$.

Consider the following countable-support prior:
\[ \Theta = \left\{ \frac{3}{4} \right\} \cup \left\{ \frac{1}{4} - \frac{1}{n^2} : n \geq 3 \right\}, \]

where distributions are indexed by the probability that they assign to outcome 1. Note that $\frac{1}{4}$ is in $\Theta$ even though it doesn’t exactly correspond to any of the agent’s conceivable models. Define $p(n) = \frac{1}{4} - \frac{1}{n^2}$.

Suppose that the agent’s utility function is given by
\[ u(a, 0) = 0 = u(b, 1), u(a, 1) = 1, u(b, 0) = 4/5, \]

Then $b$ is not preferred to $a$ for any beliefs with\textsuperscript{26}

\[ \nu(\{3/4\}) > 1/2, \]

\textsuperscript{26}Notice that the outcome distribution most favorable to action $b$ and least favorable to action $a$ is $p(3) = 1/4 - 1/9 = 5/36$. Therefore, if $\nu(\{3/4\}) > 1/2$,

\[ \sum_{n=3}^{\infty} p(n)u(a, 1)\nu(\{p(n)\}) + \frac{3}{4}u(a, 1)\nu(\{3/4\}) \geq \frac{5}{36}u(a, 1)(1 - \nu(\{3/4\})) + \frac{3}{4}u(a, 1)\nu(\{3/4\}) > 4/9 \]

and

\[ \sum_{n=3}^{\infty} (1 - p(n))u(b, 0)\nu(\{p(n)\}) + \frac{1}{4}u(b, 0)\nu(\{3/4\}) \leq \frac{31}{36}u(b, 0)(1 - \nu(\{3/4\})) + \frac{1}{4}u(b, 0)\nu(\{3/4\}) < 4/9. \]
and it is strictly preferred to a if

\[ \nu(\{3/4\}) < 1/3. \]

Then a is a Berk-Nash equilibrium but not a uniform Berk-Nash equilibrium, yet play can converge to it with positive probability from a prior \( \mu_0 \) we specify below.

Notice that

\[
1 \leq p^*(1) \left( \frac{3}{4} \frac{1}{p(n)(1)} \right)^{l_n} + p^*(0) \left( \frac{1}{4} \frac{1}{p(n)(0)} \right)^{l_n} = \\
= \frac{1}{2} \left( \frac{3}{4} \frac{1}{1 - \frac{1}{n^2}} \right)^{l_n} + \frac{1}{2} \left( \frac{1}{4} \frac{1}{\frac{3}{4} + \frac{1}{n^2}} \right)^{l_n}
\]

where

\[ l_n = \frac{\log \left( 1 - \frac{1}{n^2 + 3} \right)}{\log \left( \frac{1}{1 - \frac{1}{n^2}} \right) + \log 3} \]

Then by Dubins’ inequality, for all \( K_1 \), and \( K_2 \) there exists \( C_n \leq \frac{1}{2 \sum_{j=3}^{n^2} \frac{1}{n^2}} \) such that if \( \mu_0(p(n)) \leq C_n \) and \( \mu_0 \left( \frac{3}{4} \right) \geq \frac{1}{2} \), the probability that \( \limsup_t \frac{\mu_t(p(n))}{\mu_t \left( \frac{3}{4} \right)} > \frac{1}{n^2} K_1 \) is smaller then \( \frac{1}{n^2} K_2 \). Let \( \mu_0(p(n)) = C_n \) and \( \mu_0 \left( \frac{3}{4} \right) = 1 - \sum_{n=3}^{\infty} C_n > \frac{1}{2}, K_2 < \frac{1}{2 \sum_{n=3}^{\infty} \frac{1}{n^2}} \) and \( K_1 < \frac{1}{2 \sum_{n=3}^{\infty} \frac{1}{n^2}} \).

By the union bound with probability

\[ 1 - K_2 \sum_{n=3}^{\infty} \frac{1}{n^2} > 0 \]

we have that

\[
\limsup_t \frac{\sum_{n=3}^{\infty} \mu_t(p(n))}{\mu_t \left( \frac{3}{4} \right)} \leq \sum_{n=3}^{\infty} \limsup_t \frac{\mu_t(p(n))}{\mu_t \left( \frac{3}{4} \right)} \leq K_1 \sum_{n=3}^{\infty} \frac{1}{n^2} < \frac{1}{2}.
\]

\(^{27}\)If \( \nu_1(\{3/4\}) < 1/3 \),

\[
\sum_{n=3}^{\infty} p(n)a(1,1)\nu(p(n)) + \frac{3}{4} u(a,1)\nu(\{3/4\}) \leq \frac{1}{4} u(a,1)(1 - \nu(\{3/4\})) + \frac{3}{4} u(a,1)\nu(\{3/4\}) < \frac{5}{12} 
\]

and

\[
\sum_{n=3}^{\infty} (1 - p(n))a(b,0)\nu(p(n)) + \frac{1}{4} u(b,0)\nu(\{3/4\}) \geq \frac{3}{4} u(b,0)(1 - \nu(\{3/4\})) + \frac{1}{4} u(b,0)\nu(\{3/4\}) = \frac{7}{15}.
\]

37
Proof of Proposition 2. Consider a uniformly strict Berk-Nash equilibrium \( a \), an optimal policy \( \pi \) and \( \kappa \in (0, 1) \). By Lemma 1, there exists an \( \varepsilon \) such that if \( \nu(\hat{D}_{\varepsilon,a}) \geq 1 - \varepsilon \), then \( \pi(\nu) = a \).

For every \( l \in (0, 1) \), define the function \( f_l : P \times P \rightarrow \mathbb{R} \) by

\[
f_l(\bar{p}, p') = \sum_{y \in Y} p^*_a(y) \left( \frac{p_a(y)}{p'_a(y)} \right)^l.
\]

By Lemma 3 and since \( f_l \) is lower semicontinuous in its first argument, there exists \( \varepsilon_1 \) such that \( \bar{\nu} \in M_{l,\mu_0,a} (p^*_a) \) implies that \( f_l(\bar{p}, p') \geq 1 \) for all \( p' \) with \( p' \notin M_{l,\mu_0,a} (p^*_a) \). Let \( K = (\frac{\varepsilon}{1 - \varepsilon})^l \). Then

\[
\left( \frac{1 - \nu(\hat{D}_{\varepsilon,a})}{\nu(\hat{D}_{\varepsilon,a})} \right)^l < K \iff \left( \frac{1 - \nu(\hat{D}_{\varepsilon,a})}{\nu(\hat{D}_{\varepsilon,a})} \right)_l < \frac{\varepsilon}{1 - \varepsilon}
\]

\[

\Rightarrow \nu(\hat{D}_{\varepsilon,a}) > 1 - \varepsilon \Rightarrow \pi(\nu) = a.
\]

Let \( \bar{\varepsilon} \) be such that \( \mu'_0 \in B_\varepsilon(\nu') \) implies that

\[
\left( \frac{1 - \mu'_0(\hat{D}_{\varepsilon,a})}{\mu'_0(\hat{D}_{\varepsilon,a})} \right)^l < K (1 - \kappa).
\]

Then, \( A(\mu'_0) = \{a\} \) and combining Lemma 2, Dubins’ upcrossing lemma, and the union bound, there is a probability \( (1 - \kappa) \) that the positive supermartingale

\[
\left( \frac{1 - \mu'_t(\hat{D}_{\varepsilon,a})}{\mu'_t(\hat{D}_{\varepsilon,a})} \right)^l
\]

never rises above \( K \), so the action played is always \( a \).

Proof of Proposition 3. Let \( \nu \in \Delta(D_n) \) and fix any \( \varepsilon > 0 \). We construct an initial belief \( \nu_\varepsilon \) that is \( \varepsilon \) close to \( \nu \) but such that the action process never converges to \( a \). The first step is to observe that since \( a \) is not a uniform Berk-Nash equilibrium, there is KL-minimizer \( \hat{p} \) for which \( a \) is not myopically optimal. Next, we associate each outcome distribution in the support of \( \nu \) with an outcome distribution that is \( \varepsilon \) close to it but that it is not a KL minimizer, and we define \( \bar{\nu} \) on the set of associated distributions. Finally, \( \nu_\varepsilon \) is defined as
the linear combination between the Dirac measure on $\hat{p}$ and $\bar{\nu}$. Therefore, the unique KL-minimizer in the support of $\nu_\varepsilon$ is $\hat{p}$, and we can use Proposition 1 to rule out convergence to $a$.

Since $a$ is not a uniform Berk-Nash equilibrium, there exists $\hat{p} \in D_a$ with $a \notin A_m(\delta_\hat{p})$. For every $p \in D_a$, choose $q_{p,\varepsilon} \in B_\varepsilon(p) \setminus D_a$, and $\bar{\varepsilon}_{p,\varepsilon}$ be such that $||q' - q_{p,\varepsilon}|| < \bar{\varepsilon}_{p,\varepsilon}$ implies $q' \notin \text{argmin}_{p \in \text{supp} \nu} H(p_a^*, p_a')$.

Define $\Phi_\varepsilon : D_a \to 2^\Theta$ as

$$\Phi_\varepsilon(p) = \{ p' \in B_\varepsilon(p) \cap B_{\varepsilon_{p,\varepsilon}}(q_{p,\varepsilon}) \}.$$ 

The correspondence $\Phi_\varepsilon$ is nonempty and closed valued, so it has a measurable selection $\phi_\varepsilon$ by the Kuratowski Selection Theorem (see, e.g., Theorem 18.13 in Aliprantis and Border, 2013).

Define $\nu_\varepsilon = (1 - \varepsilon) \delta_{\hat{p}} + \varepsilon \bar{\nu}_\varepsilon$ where

$$\bar{\nu}_\varepsilon(\psi) = \nu(\phi_\varepsilon^{-1}(\psi)).$$

Then $\nu_\varepsilon \to \nu$, but argmin $H(p_a^*, p_a') = \{ \hat{p} \}$, so by Proposition 1, the probability of converging to $a$ starting from belief $\nu_\varepsilon$ is 0.

**Proof of Proposition 4.** Consider a uniformly strict Berk-Nash equilibrium $a$. By the generalization of Weierstrass Theorem to lower-semicontinuous functions (see, e.g., Aliprantis and Border, 2013), $D_a$ is a compact set. Therefore, $\Delta(D_a)$ is compact (see, e.g., Theorem 6.4 in Parthasarathy, 2005).

To ease notation, in this proof for every $\varepsilon > 0$ and $q \in \Delta(Y)$ we let $M_\varepsilon(q) = M_{\varepsilon, \mu_{0,a}}(q)$. Denote as $\overline{M}_\varepsilon(p_a^*)$ the closure of $M_\varepsilon(p_a^*)$, and define the function

$$G(\varepsilon) = \min_{\nu \in \Delta(\overline{M}_\varepsilon(p_a^*))} \left( V_\nu(a) - \max_{a' \neq a} V_\nu(a') \right).$$

By the Maximum Theorem, $G$ is continuous, and there exists $\varepsilon' > 0$ such that if $\varepsilon' > \varepsilon$, $G(\varepsilon) > 0$, that is, $\nu \in \Delta(\overline{M}_\varepsilon(p_a^*))$ implies $A_m(\nu) = \{a\}$.

Now define the function $f_l$ by

$$f_l(\bar{p}, p') = \sum_{y \in Y} p_a^*(y) \left( \frac{\bar{p}(y)}{p'(y)} \right)^l.$$
As in Frick, Iijima, and Ishii (2019b), a simple continuity argument shows that for each $\bar{p} \in \text{argmin} H(p_a^*, p)$ and $p' \notin M_\varepsilon(p_a^*)$ there exists an $l(\bar{p}, p')$ such that $f_l(\bar{p}, p') > 1$ for all $l \leq l(\bar{p}, p')$.

Suppose by way of contradiction that there was no $l \in (0, 1)$ such that $f_l(\bar{p}, p') > 1$ for all $\bar{p} \in \text{argmin} H(p_a^*, p)$ and $p' \notin M_\varepsilon(p_a^*)$. Then, we can define a sequence $(\bar{p}_n, p'_n)$ such that $f_{\frac{1}{n}}(\bar{p}_n, p'_n) \leq 1$, and sequential compactness of $\text{argmin} H(p_a^*, p) \times M_\varepsilon(p_a^*)$ guarantees that this sequence has an accumulation point $(\bar{p}, p')$. However, for, $n > \frac{1}{l(\bar{p}, p')}$, $f_{\frac{1}{n}}(\bar{p}n, p'n) \leq 1$ implies $f_l(\bar{p}, p')(\bar{p}_n, p'_n) \leq 1$, but then continuity of $f_l(\bar{p}, p')$ leads to a contradiction with $f_l(\bar{p}, p') > 1$. Since $f_l$ is continuous, there exists $\varepsilon' \in (0, \varepsilon)$ such that $\bar{p} \in M_\varepsilon(p_a^*)$ implies $f_l(\bar{p}, p') \geq 1$ for all $p' \notin M_\varepsilon(p_a^*)$. Let $K$ be such that $\frac{\mu'_{\varepsilon'(M_{\varepsilon'}(p_a^*))}}{1 - \mu'_{\varepsilon(M_{\varepsilon}(p_a^*))}} > K$ implies $A(\mu'_{\varepsilon'}) = a$.

Using the Maximum Theorem again we can find a sequence of outcome realizations $y^t$ such that if $\hat{p}_t$ is the corresponding empirical frequency, it is sufficiently close to $p_a^*$ to have

$$M_{\varepsilon'/2}(\hat{p}_t) \subseteq M_{\varepsilon'}(p_a^*).$$

Therefore by Lemma 4, there exists a time period $T$ such that for all $t' > T$, if the empirical frequency $\hat{p}_{t'} = \hat{p}_t$, the agent assigns a relative probability higher than $K$ to an $\varepsilon'$ Ball around $\bar{p}$

$$\frac{\mu'_{\varepsilon'}(M_{\varepsilon'}(p_a^*))}{1 - \mu'_{\varepsilon'}(M_{\varepsilon}(p_a^*))} > K.$$

Notice by replicating the outcome realizations $y^t$ sufficiently many time, we have a sequence of outcomes $y^{t'}$ such that the empirical frequency $\hat{p}_{t'} = \hat{p}_t$ and $t' > T$. Since $p_a^*$ is absolutely continuous with respect to to $p_{a'}^*$ for all $a' \in A$, the previous sequence of outcomes has positive probability, and after this outcome sequence the agent plays $a$. By Lemma 2 and the law of iterated expectations, conditional on $a$ being played $(\frac{1 - \mu'_{\varepsilon}(M_{\varepsilon'/2}(\hat{p}_t))}{\mu'_{\varepsilon'}(M_{\varepsilon'}(p_a^*))})^l$ is a positive supermartingale.

Then, by Dubins’ upcrossing lemma, there is a positive probability that this positive supermartingale never rises above $1/K^l$, and therefore $a$ is never stopped to be played. ■

Proof of Proposition 5. Let $b$ be a uniformly strict Berk-Nash equilibrium supported by the belief $\nu$, and to ease notation, let $\mu$ denote the prior throughout this proof. To focus on the additional difficulties involved in the bandit case, we prove the result in the case in which

$$\arg\min_{p_0: p \in \Theta} H(p_0^*, p) = \arg\min_{p_0: p \in \Theta} H(p_a^*, p).$$
is a singleton. The extension to multiple minimizers involves some additional continuity-compactness argument that mimic the ones in Proposition 2. Without loss of generality, we can take \( \nu \) to be such that for all \( a \in A \setminus \{b\}, \nu_a = \delta_{p(a)} \) for some \( p(a) \in \Delta(Y) \) such that \( p^*_a(y) > 0 \) implies \( p(a)(y) > 0 \). Assume for all \( a \in A \) and \( y \in Y \) such that \( p(a)(y) \) is rational, the general case with \( p(a)(y) \) possibly irrational involving a tedious additional continuity argument.

Let \( \{y(b)_{i=1}^{N_b}\} \) be a sequence of outcomes such that \( \hat{p}_{N_b}(y^{N_b}) = p(b) \), and let \( \{y(b)_{i=1}^{\infty}\} \) be its infinite repetition. By Lemma 4, for every \( \varepsilon > 0 \), there exists \( K_\varepsilon \) such that for all \( t > K_\varepsilon, \mu_b(B_{\varepsilon}(p_b) \mid y(b)^t) > 1 - \varepsilon_b \).

Let \( \beta \in (0, 1) \) and \( (\varepsilon_a)_{a \in A} \in \mathbb{R}_+^A \) be such that if \( \beta > \beta \) and the belief \( \hat{\nu} \) is such that \( \hat{\nu}_b \in \{\mu_b (\cdot \mid y(b)^t)\}_{t=1}^{K_\varepsilon} \bigcup \{\nu'_b : \nu'_b(B_{\varepsilon_b}(p_b)) > 1 - \varepsilon_b\} \), and for all \( a' \neq b \), \( \nu(a')(B_{\varepsilon_b}(p_b)) > 1 - \varepsilon_{a'} \) then the highest Gittins index is the one of action \( b \). Their existence is guaranteed by \( \{a\} = A^m(\nu) \) and the definition of Gittins index. For each \( \beta > \beta \), let \( \varepsilon_\beta < \varepsilon_b \) be such that if \( \nu_b(B_{\varepsilon_\beta}(p_b)) > (1 - \varepsilon_\beta) \) then the probability of converging to play action \( a \) is larger than \( \frac{1}{2} \) under any optimal policy given the discount factor \( \beta \), whose existence is guaranteed by Proposition 2.

By Lemma 4 for every \( a \neq b \) there exists a finite number \( n_a \) such that after \( n_a \) observations \( \nu_a(B_{\varepsilon_a}(p_a) \mid \hat{p}_{n_a} = p(a)) > 1 - \varepsilon_a \). For every \( a \neq b \), let \( \bar{n}_a \geq n_a \) and \( \{y(a)_{i=1}^{n_a}\} \) be a sequence of realizations such that \( \hat{p}_{n_a}(a) = p(a) \) and \( n_a \). Finally, let \( n_b = K_{\varepsilon_\beta} \). Then, the array \( \{\{y(a)_{i=1}^{n_a}\}_{a \in A}\} \) has positive probability, the agent starts to play \( a \) after at most \( \sum_{a \in A} n_a \) periods, and with probability \( \frac{1}{2} \) continues to play \( a \) forever.

**Proof of Proposition 7.** We prove the statement for \( \bar{a}, \) the other is analogous. Denote the optimal policy used by the agent as \( \pi \). Since the environment is strongly supermodular and \( \bar{a} \) is a uniformly strict Berk-Nash equilibrium, by Proposition 2 there exists \( \bar{p} \in \mathcal{M} \) and \( K \in (0, 1) \) such that if \( \nu(\{p : p > \bar{p}\}) > K \), then the probability that \( a \) is going to be used forever is larger than \( \frac{1}{2} \). Denote the highest outcome as \( \bar{y} \). Since the environment is strongly supermodular, for every action \( b \in A \),

\[
\frac{\mu_{t+1} \{(p : p > \bar{p}) \mid (a^t, y^t), (b, \bar{y})\}}{1 - \mu_{t+1} \{(p : p > \bar{p}) \mid (a^t, y^t), (b, \bar{y})\}} > \frac{\mu_t \{(p : p > \bar{p}) \mid (a^t, y^t)\}}{1 - \mu_t \{(p : p > \bar{p}) \mid (a^t, y^t)\}}.
\]

Therefore, there exists a finite number \( n(b) \) such that if \( a_t = b \) and \( y_t = \bar{y} \) for all \( t \leq n(b) \), then

\[
\mu_t \{(p : p > \bar{p}) \mid (a^t, y^t)\} \geq K.
\]

Consider the event \( E \) that for all \( b \in A \) and \( t \leq n(b) \), \( x_{t,b} = \bar{y} \). This event has strictly
After having observed signals, the choice would not affect our results.

Let \( p \) be the probability space of action-dependent outcome realizations, a Berk-Nash equilibrium.

We start by specifying the details of the sample space \( \Omega = S \times (Y^\infty)^A \). Formally, we consider the probability space \( \Omega, \mathcal{F}, \mathbb{P} \), where \( \mathcal{F} \) is the discrete sigma algebra and the probability measure \( \mathbb{P} \) is the product measure induced by independent draws (across signal, actions, and time) according to \( p^* \).

We denote the outcome observed by the agent in period \( t \) after action \( a_t \) by \( y_t = x_{k,a_t} \), where \( k = \{|s \leq t: a_s = a_t|\} \) is the number of times the agent has taken action \( a_t \) has taken up to and including period \( t \). A (pure) policy \( \pi : \bigcup_{t=0}^\infty S^{t+1} \times A^t \times Y^t \rightarrow A \) specifies an action for every history \((s_1, a_1, y_1, s_2, a_2, y_2, \ldots, s_t, a_t, y_t, s_{t+1})\), and an initial action \( a_1 \). Throughout, we denote by \( a_{t+1} = \pi(s^{t+1}, a^t, y^t) \) the action taken in period \( t \) where \((s^{t+1}, a^t, y^t)\) is a sequence of realized signals, actions, and outcomes.\(^{28}\)

**Proof of Proposition 8**  Since there is a finite number of actions, if behavior converges to \( \sigma \), there exists a finite period \( T \) such that \( a_t((s_k, x_{a,s',k})_{k \in \mathbb{N}, a \in A, s' \in S}) = \sigma(s_t) \) for all later time periods \( t > T \). Consider an arbitrary period \( T \in \mathbb{N} \), strategy \( \sigma \in A^S \), and initial sequence of signals, actions, and outcomes \((s^T, a^T, y^T)\). We will show that almost surely, if \( a_t((s_k, x_{a,s',k})_{k \in \mathbb{N}, a \in A, s' \in S}) = \sigma(s_t) \) for all periods \( t \) after the initial sequence \((s^T, a^T, y^T)\), \( \sigma \) is a Berk-Nash equilibrium.

The initial sequence \((s^T, a^T, y^T)\) corresponds to the event

\[
E = \{ (s_k, x_{a,s',k})_{k \in \mathbb{N}, a \in A, s' \in S} : \forall t \leq T, y_t = x_{t|\{\tau \leq t: a_\tau = a_t\}|,s_t,a_t} \}.
\]

Let \( K_s = |\{\tau \leq T : (s_\tau, a_\tau) = (s, \sigma(s))\}| \) be the number of times action \( \sigma(s) \) has been played after having observed signal \( s \) up to and including period \( T \).

\(^{28}\)Multiple state spaces lead to the same law for the stochastic processes we are interested in. In particular, we could have started from the probability space of action-dependent outcome realizations \((s_k, x_{a,s',k})_{k \in \mathbb{N}, a \in A, s' \in S}\), but with \( x_{a,s,k} \) denoting the outcome realization if signal \( s \) realizes and agent takes action \( a \) at period \( k \). An argument similar to that of Lemma 4 of Fudenberg and He (2017) shows that this choice would not affect our results.
Next, we define an ancillary stochastic process on our probability space, \((\nu_{t,\sigma})_{t \geq T}\). The stochastic process is recursively defined as follows:

\[
\nu_{T,\sigma} = \mu_0 \left( \{ (s^T, a^T, y^T) \} \right)
\]

\[
\nu_{t+1,\sigma}(\Psi) = \frac{\int_{\Psi} p_{s_{t+1},\sigma(s_{t+1})}(x_{\sigma(s_{t+1}),s_{t+1}+1},N_{s_{t+1}+1,t}) \, d\nu_{t,\sigma}(p)}{\int_{\Psi} p_{s_{t+1},\sigma(s_{t+1})}(x_{\sigma(s_{t+1}),s_{t+1}+1},N_{s_{t+1}+1,t}) \, d\nu_{t,\sigma}(p)},
\]

where \(N_{s,t} = K_s + \{ \{ \tau: T < \tau \leq t: s_\tau = s \} \}\). Here \(\nu_{t,\sigma}\) would be the agent’s period \(t\) posterior belief after having observed the initial sequence \((s^T, a^T, y^T)\) and having followed strategy \(\sigma\) in all periods \(\tau \in (T,t]\). In general, the process \((\nu_{t,\sigma})_{t \geq T}\) will not coincide the evolution of the agent’s beliefs \((\mu_t)_{t \geq T}\): they may choose an action different from the one prescribed by strategy \(\sigma\), leading to different observations and posteriors.

We will prove a statement about \((\nu_{t,\sigma})_{t \geq T}\) that holds almost everywhere in \(E\), and then we argue that \(E\) can partitioned in two subsets, the sequences along which the agent stops to follow \(\sigma\) at some period \(t > T\), and the sequences along which \(\sigma\) is played forever. In the former subset, clearly we do not have convergence to \(\sigma\) after \((s^T, a^T, y^T)\). In the latter, the posterior belief \((\mu_t)_{t \geq T}\) will coincide with \((\nu_{t,\sigma})_{t \geq T}\), and therefore it will inherit its properties.\(^{29}\)

**Claim 1.** For every neighborhood \(\Psi\) of \(\arg\min_{p \in \text{supp} \mu_T} \sum_{s \in S} \zeta(s) H \left( p_{\sigma(s),s}^*, p_{\sigma(s),s} \right) \),

\[
P \left[ \lim_{k \to \infty} \nu_{k,\sigma}(\Psi) = 1 \mid E \right] = 1.
\]

**Proof of Claim** \(^{[4]}\). To prove the claim, notice that the probability of the sequence \((s_{T+k}, x_{\sigma(s'),s'}, K_{s'+k})_{k \in \mathbb{N}, s' \in S}\) conditional on \(E\) is the product measure induced by independent draws. Let \(Z = S \times Y\).

If the strategy of the agent were exogenously fixed at \(\sigma\) for every period \(t \geq T\), the signal and outcome pair \((s_t, y_t)\) would be i.i.d with probability \(p_{\sigma}^* (s, y) : = \zeta(s) p_{\sigma(s), s}^* (y)\), whereas \(p \in \text{supp} \mu_T\) would assign \(p_{\sigma}(s, y) : = \zeta(s) p_{s,\sigma(s)}(y)\).

Since (1) \(\text{supp} \mu_T\) is a compact set, (2) \(\mu_T\) is a Borel probability measure on \(\text{supp} \mu_T\), and assumptions (ii), (iii), and (iv) of Berk [1966] are satisfied given the absolute continuity assumption and the finite number of outcomes, it follows from Berk’s Theorem that there

\(^{29}\)Notice that the evolution of \(\nu_{t,\sigma}\) only depends on the sequences \((s_{T+k}, x_{\sigma(s'),s'}, K_{s'+k})_{k \in \mathbb{N}, s' \in S}\).
exists $E'$ with $\mathbb{P}[E' \mid E] = 1$ such that for every neighborhood $Ψ$ of

$$\arg\min_{p \in \text{supp} \mu} H \left( p^*_{\sigma}, p \right)$$

$$= \arg\min_{p \in \text{supp} \mu} \left( - \sum_{(s,y) \in S \times Y} \zeta(s) p^*_{\sigma(s)}(y) \log \zeta(s) p_{\sigma(s)}(y) \right)$$

$$= \arg\min_{p \in \text{supp} \mu} \left( - \sum_{(s,y) \in S \times Y} \zeta(s) p^*_{\sigma(s)}(y) \log \zeta(s) - \sum_{(s,y) \in S \times Y} \zeta(s) p^*_{\sigma(s)}(y) \log p_{\sigma(s)}(y) \right)$$

$$= \arg\min_{p \in \text{supp} \mu} \sum_{s \in S} \zeta(s) H \left( p^*_{\sigma(s)}, p_{\sigma(s)} \right)$$

we have $\lim_{k \to \infty} \nu_{k,\sigma}(Ψ) = 1$ on $E'$.

Given our absolute continuity hypothesis, $\text{supp} \mu_T = \Theta$, and this in turns implies that

$$\arg\min_{p \in \text{supp} \mu_T} \sum_{s \in S} \zeta(s) H \left( p^*_{\sigma(s)}, p_{\sigma(s)} \right) = D_\sigma.$$

This proves that if the agent starts to follow $\sigma$ from $(s^T, a^T, y^T)$, then

$$1 = \lim_{k \to \infty} \nu_{k,\sigma}(Ψ) = \mu_t(Ψ) \text{ a.s.}$$

for every neighborhood $Ψ$ of $D_\sigma$.

By the 0-1 Law, $E'$ can be chosen such that if $(s_{T+k}, x_{\sigma(s')}, s', K_{s'+k})_{k \in \mathbb{N}, s' \in S} \in E'$, for all $s \in S$, $s_t = s$ infinitely many times. Suppose that $\sigma$ is not a Berk-Nash equilibrium. Since $P$ is compact, $\Delta(P)$ is also compact. Moreover, since $P$ is separable, $\Delta(P)$ is metrizable by the Levy-Prokhorov metric and thus sequentially compact. Hence the sequence $(\nu_{k,\sigma})_{t \in \mathbb{N}}$ has an accumulation point $\nu \in \Delta(D_\sigma)$. Since $\sigma$ is not a Berk-Nash, there exists $s \in S$ such that $\sigma(s) \notin A^m(\nu, s)$, and from the definition of weak convergence of measure and Berk-Nash equilibrium, there exists $\tau'$ such that for $\tau \geq \tau'$

$$\int_{\Delta(P)} \mathbb{E}_{p_{s,\sigma(s)}} [u(s, \sigma(s), y)] d\nu_{\tau,\sigma}(p) < \max_{a \in A} \int_{\Delta(P)} \mathbb{E}_{p_{s,a}} [u(s, a, y)] d\nu_{\tau,\sigma}(p).$$

Since $s_t = s$ infinitely often on $E'$, $s_\tau = s$ for some $\tau \geq \tau'$, and the agent will play an action $a_s = \sigma(s)$. This contradicts convergence to $\sigma$ after $(s^T, a^T, y^T)$. Since both $\sigma$ and $(s^T, a^T, y^T)$ were chosen arbitrarily, the result holds.
Proof of Proposition[9]. Under causation and signal neglect, $\Theta \subseteq \Delta (\Delta (Y))$. Consider a uniformly strict Berk-Nash equilibrium $\sigma$. By the Maximum Theorem, $D_\sigma$ is a compact set, and this implies that $\Delta(D_\sigma)$ is compact. Since $\sigma$ is the unique optimal best reply strategy at the beliefs in $\Delta(D_\sigma)$, there exists $\varepsilon \geq 0$ such that if

$$\nu \left( M_{\varepsilon} (\bar{p}_\sigma) \right) \geq (1 - \varepsilon)$$

then the myopic best reply to $\nu$ is $\sigma$. By the same argument of the proof of Proposition 2 there exists an $l \in (0, 1)$ and $\varepsilon' \in (0, \varepsilon)$, such that if $p \in M_{\varepsilon'} (\bar{p}_\sigma)$ and $p' \notin M_{\varepsilon} (\bar{p}_\sigma)$ then $f_l(p, p') \geq 1$.

Using the Maximum Theorem again we can find a sequence of outcome realizations $y_t$ such that if $\hat{p}_t$ is the corresponding empirical frequency, it is sufficiently close to $\bar{p}_\sigma$ to have $M_{\varepsilon/2} (\hat{p}_t) \subseteq M_{\varepsilon} (\bar{p}_\sigma)$.

Therefore by Lemma 4, there exists a time period $T$ such that for all $t' > T$, if the empirical frequency $\hat{p}_{t'} = \hat{p}_t$, the agent assigns a relative probability higher than $K$ to an $\varepsilon$ Ball around $\bar{p}$

$$\frac{\mu_{t'}(M_{\varepsilon} (\bar{p}_\sigma))}{1 - \mu_{t'}(M_{\varepsilon} (\bar{p}_\sigma))} \geq \frac{\mu_{t'}(M_{\varepsilon/2} (\bar{p}_\sigma))}{1 - \mu_{t'}(M_{\varepsilon} (\bar{p}_\sigma))} > 2 \frac{(1 - \varepsilon)}{\varepsilon}.$$  

Notice by replicating the outcome realizations $y'_t$ sufficiently many time, we have a sequence of outcomes $y_{t'}$ such that the empirical frequency $\hat{p}_{t'} = \hat{p}_t$ and $t' > T$. Since $\text{supp} p^*_{a, s} = Y$ for all $(a, s) \in A \times S$, the previous sequence of outcomes has positive probability, and after this outcome sequence the agent plays $\sigma$. By Lemma 2 and the law of iterated expectations, conditional on $a$ being played

$$\left( \frac{1 - \mu_{t'}(M_{\varepsilon} (\bar{p}_\sigma))}{\mu_{t'}(M_{\varepsilon} (\bar{p}_\sigma))} \right)^l$$

is a positive supermartingale.

Then, by Dubins’ upcrossing lemma, there is a positive probability that this positive supermartingale never rises above $\frac{\varepsilon}{(1 - \varepsilon)}$, that in turns imply that $\mu_{t'}(M_{\varepsilon'/2} (\hat{p}_t))$ never goes below $(1 - \varepsilon)$ and therefore $\sigma$ is always played after the sequence $y'$.  

### 8.4 Uniform version of Diaconis and Freedman

For every $\varepsilon > 0$, $K > 0$, we say that a probability measure $\nu \in \Delta(\Delta(Y))$ is $(\varepsilon, K)$-dense if it assigns probability greater $K$ to every $\varepsilon$-ball, i.e. $\nu(B_\varepsilon(p)) > K$ for every $p \in \Delta(Y)$.  

45
Lemma 5. For all $\varepsilon \in (0,1)$ and $h \in (0,\frac{1}{4})$ such that $\mu_0$ is $(\varepsilon h^2, K)$-dense,
\[
\frac{\mu_t (B_h(\hat{p}_t))}{1 - \mu_t (B_h(\hat{p}_t))} \geq Ke^{t(1-\varepsilon)2h^2}.
\]

Proof of Lemma 5. In this setting, denote as $\hat{p}_t = \frac{1}{t} \sum_{k=1}^{t} y_k$ the fraction of 1 signals. Bayesian updating gives that the subjective belief of the agent after observing a sequence of signals where the frequency of 1s equals $\hat{p}_t$ is given by
\[
\frac{\mu_t ([\hat{p}_t - h, \hat{p}_t + h])}{1 - \mu_t ([\hat{p}_t - h, \hat{p}_t + h])} = \int_{[\hat{p}_t-h, \hat{p}_t+h]} e^{-tH(\hat{p}_t, \sigma)} d\mu_0(\sigma).
\]

The denominator can be bounded by
\[
\int_{[0, \hat{p}_t-h] \cup [\hat{p}_t+h, 1]} e^{-tH(\hat{p}_t, \sigma)} d\mu_0(\sigma) \leq \sup_{\sigma \in [0, \hat{p}_t-h] \cup [\hat{p}_t+h, 1]} e^{-tH(\hat{p}_t, \sigma)}. \tag{8}
\]

Since $\sigma \mapsto H(\hat{p}_t, \sigma)$ is convex, $\sigma \mapsto e^{-tH(\hat{p}_t, \sigma)}$ is quasi-concave. As $\sigma \mapsto H(\hat{p}_t, \sigma)$ has a strict minimum at $\hat{p}_t$ it thus follows that $\sigma \mapsto e^{-tH(\hat{p}_t, \sigma)}$ is non-decreasing for $\sigma < \hat{p}_t$ and non-increasing for $\sigma > \hat{p}_t$. This implies that (8) can be bounded by
\[
\max \{ e^{-tH(\hat{p}_t, \hat{p}_t-h)}, e^{-tH(\hat{p}_t, \hat{p}_t-h)} \} \leq e^{-t[H(\hat{p}_t, \hat{p}_t-h)+\inf_{p, \sigma} \{H(p, \sigma)-H(p, \hat{p}_t)|\sigma-p|>\sigma]}..\]

where the second inequality follows from Corollary 3.5 of Diaconis and Freedman [1990].

For the numerator suppose wlog $\hat{p}_t \in [0,\frac{1}{2}]$. Take the derivative of the relative entropy:
\[
\frac{\partial H(\hat{p}_t, \sigma)}{\partial \sigma} = -\frac{\hat{p}_t}{\sigma} + \frac{(1-\hat{p}_t)}{1-\sigma}.
\]

When evaluated at $\sigma \in (\hat{p}_t, \hat{p}_t+h)$, the first term ranges between $-1$ and $0$, the second term
between 1 and \( \frac{(1-p_t)}{(1-p_t-h)} < \frac{1/2}{1/2-1/4} = 2 \) and therefore

\[
\left| \frac{\partial H (\hat{p}_t, \sigma)}{\partial \sigma} \right| < 2.
\]  
(9)

By assumption, \( K \leq \mu_0 ([\hat{p}_t, \hat{p}_t + \varepsilon h^2]) \). By (9) if \( \sigma \in [\hat{p}_t, \hat{p}_t + \varepsilon h^2] \),

\[
H (\hat{p}_t, \sigma) \leq H (\hat{p}_t, \hat{p}_t) + 2\varepsilon h^2.
\]

Therefore, as \( \hat{p}_t + h \geq \hat{p}_t + \varepsilon h^2 \), footnoteThis is true as by assumption \( \varepsilon < 1, h < \frac{1}{4}. \)

\[
\int_{[\hat{p}_t-h, \hat{p}_t+h]} e^{-tH(\hat{p}_t, \sigma)}d\mu_0(\sigma) \geq \int_{[\hat{p}_t, \hat{p}_t+\varepsilon h^2]} e^{-tH(\hat{p}_t, \sigma)}d\mu_0(\sigma) \\
\geq e^{-t(H(\hat{p}_t, \hat{p}_t)+2\varepsilon h^2)}K.
\]

Combining the two bounds we get the result. ■

References


48


