Limit Points of Endogenous Misspecified Learning[*]

Drew Fudenberg[†] Giacomo Lanzani[‡] Philipp Strack[§]

First posted version: March 10, 2020

This version: December 15, 2020

Abstract

We study how a misspecified agent learns from endogenous data when their prior belief can assign probability 0 to a neighborhood of the true model. We show that only *uniform Berk-Nash equilibria* can be long-run outcomes, and that all *uniformly strict Berk-Nash equilibria* have an arbitrarily high probability of being the long-run outcome for some initial beliefs. When the agent believes the outcome distribution is exogenous, every uniformly strict Berk-Nash equilibrium has positive probability of being the long-run outcome for any initial belief. We generalize these results to settings where the agent observes a signal before acting.

Keywords: Misspecified learning, Bayesian consistency, Berk-Nash equilibrium

[*] We thank Aislinn Bohren, Alex Wolitzky, Annie Liang, Ben Golub, Cuimin Ba, Demian Pouzo, Glenn Ellison, Ignacio Esponda, Harry Pei, Kevin He, Mira Frick, Roberto Corrao, Ryota Iijima, Yuhta Ishii, and Yuichi Yamamato for helpful conversations, and National Science Foundation grants SES 1643517 and 1951056, the Guido Cazzavillan Scholarship, and Sloan foundation 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 data generating process. Given the complexity of the real world, such misspecification is plausible in many settings, and has been studied in a wide range of applications.

We consider a general environment with finite actions and outcomes and – unlike most past work – do not restrict the agent's prior belief to have a finite support or any specific functional form. In this 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 map from actions to distribution over outcomes. Especially, the agent's prior 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, Theorem 1 shows that regardless of the agent's discount factor, if play converges to an action a, that action is a *uniform Berk-Nash equilibrium.* Uniform Berk-Nash equilibrium, which we introduce in this paper, is a refinement of Berk-Nash equilibrium (Esponda and Pouzo, 2016). Berk-Nash equilibrium requires that the action is myopically optimal against some belief that minimizes the Kullback-Leibler (KL) divergence between the subjective and true outcome distributions given that the agent plays a— that is, a best response to a "KL minimizer". Uniform Berk-Nash equilibrium strengthens this by requiring that the action is a best response to *any* beliefs with support on these KL minimizers. Intuitively, limit points correspond to myopic optimization even when the agent is not myopic because play will not converge until the agent no longer perceives an "experimentation value" from non-myopic play; the intuition for the uniformity requirement is that when play converges, the agent's beliefs oscillate over all of the KL-minimizing beliefs.

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 arbitrarily high probability for some open set of initial beliefs. Theorem 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 KL-minimizing outcome distributions (as opposed to being true for at least one of them).

We say that an action is *positively attractive* if there is positive probability that it is the limit outcome under every optimal policy for *every* full-support prior belief. When the agent believes (either rightly or wrongly) that the distribution of outcomes is "exogenous", i.e. the same for all actions, or in a "subjective bandit" problem, where the agent believes that the outcomes observed when playing one action are uninformative about the outcome distributions induced by other actions, we obtain partial converses to Theorem 1: All uniformly strict Berk-Nash equilibria are positively attractive. 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 prove in Proposition 1 that beliefs concentrate exponentially fast around the KL minimizers.[1] We use Proposition 1 to guarantee that the agent starts to play the equilibrium action with positive probability. We then use the stability result from Theorem 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. In this setting, 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 uniform Berk-Nash equilibrium. Moreover, if the agents ignore the predictive value of the signals, i.e., the signals are *subjectively uninformative*, every uniformly strict Berk-Nash equilibrium is positively attractive.

We illustrate our findings in three economic examples: a monopolist that is misspecified about the demand function, a central bank choosing an exchange-rate policy, and a seller that observes a signal and then decides whether to make an investment.

1.1 Related Work

Berk (1966) shows that the beliefs of a misspecified agent asymptotically concentrate on the set of models that minimize the KL divergence from the true data generating process when this process 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, so the agent's misspecification has implications for what they observe and thus for their long-run beliefs. Arrow and Green (1973) gives the first general framework

[1] This result is in the spirit of Diaconis and Freedman (1990), but it is more special as it does not hold uniformly over all empirical distributions. The result of Diaconis and Freedman does not apply, as its assumptions rule out both misspecification and finite-support priors.

for this problem, and Nyarko (1991) points out that the combination of misspecification and endogenous observations can lead to cycles.

There has been a surge of theoretical work on misspecified learning since the seminal work of Esponda and Pouzo (2016), which defines Berk–Nash equilibrium. This is a relaxation of Nash equilibrium that replaces 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 show that Berk-Nash equilibrium is a necessary property for limit points when the payoff function is subject to small i.i.d. random shocks as in Fudenberg and Kreps (1993), and that it is sufficient if in addition the agent is willing to incur asymptotically negligible optimization losses.

Fudenberg, Romanyuk, and Strack (2017) and Bohren and Hauser (2020) provide necessary and sufficient conditions for actions to converge when the support of the agent's prior contains only two points.[2] Heidhues, Kőszegi, and Strack (2018b) and He (2019) provide conditions for global convergence of play of a non-myopic agent in a environments with additively separable payoffs that satisfy strong supermodularity restrictions, where the Berk-Nash equilibrium is unique. Heidhues, Kőszegi, and Strack (2018a) establishes convergence to a Berk-Nash equilibrium in environments with a normal prior and normal signals. 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.

The most closely related papers are Esponda, Pouzo, and Yamamoto (2019) (henceforth EPY) and Frick, Iijima, and Ishii (2020) (henceforth FII). EPY uses stochastic approximation to establish when the agent's *action frequency* converges in an environment with finitely many actions and fairly general priors. We provide a sharper characterization of when play converges to a single action in the long run, but our results do not characterize the long-run distribution when this convergence does not occur. Corollary 3 in the Appendix combines our results with theirs to derive new results about the limiting action frequencies. FII provides conditions for local and global convergence of the agent's beliefs without explicitly modelling the agent's actions when the agent's prior has finite support.[3]

Our paper complements the literature on long-run behavior in misspecified models in three ways: First, we establish that without the asymptotically vanishing payoff perturbations of Esponda and Pouzo (2016), play never converges to a non-uniform Berk-Nash

[2] Bohren and Hauser (2020) considers myopic agents in discrete time; Fudenberg, Romanyuk, and Strack (2017) analyzes a continuous time model with Brownian noise without assuming myopia.

[3] Neither model nests the other. FII assumes finite priors, and impose a continuity assumption that our model can but need not satisfy. Conversely, we rule out the continuum of actions assumed by FII.

equilibrium. (This uniformity refinement has no analog in Frick, Iijima, and Ishii (2020) because it is is with respect to the optimality of actions.) Second, we introduce conditions under which an action has positive probability of being the long-run outcome from any initial belief. Finally, we provide the first necessary and sufficient conditions for the choices of forward-looking misspecified agents to converge to a myopic best reply to their beliefs.[4]

Misspecified agents are featured in work in a wide range of fields. There are many examples in behavioral economics, such as the "law of small numbers," the "hot-hand fallacy," the winner's curse, and the link between overconfidence and prejudice. (Kagel and Levin (1986), Rabin and Vayanos (2010), and Heidhues, Kőszegi, and Strack (2019).) Macroeconomists have been interested in misspecified learning both in the form of misspecified least-squares predictions as well as more sophisticated models of updating and inference.[5] In organizational economics, misspecification has been used to explain e.g. the role of corporate culture and the low rate and low number of minority inventors. In public economics, misspecification helps explain over or under reaction to changes in tax schedules. And in political economy, misspecification has been used to explain the recurrence of populism and political polarization.[6] There is also a related literature on misspecified social learning.[7]

In addition to papers that consider misspecified Bayesian agents, there is also a literature that studies the long-run outcomes under learning heuristics that might be used when people are unable to formulate a probabilistic assessment of the data generating process. Many of these heuristics feature a form of neglect of the 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).

2 The Model

2.1 Setup

Actions, Utilities and Objective Outcome Distributions In each period $t \in \{1, 2, 3, \ldots\}$ an agent chooses an action from the finite set A. This choice has two effects. First, each ac-

[4] Theorem 4 of Esponda and Pouzo (2016) shows that Berk-Nash is necessary under weak identification and payoff perturbations. Other work either assumes myopia or don't obtain convergence to myopic best reply.

[5] Bray (1982), Bray and Savin (1986), Cho and Kasa (2015), Cho and Kasa (2017), Molavi (2019).

[6] See Gibbons, LiCalzi, and Warglien (2019) and Bell et al. (2019) for organizational economics, Rees-Jones and Taubinsky (2016) and Morrison and Taubinsky (2019) for public economics, and Levy, Razin, and Young (2020) and Eliaz and Spiegler (2018) for political economy.

[7] E.g. Bohren (2016), Bohren and Hauser (2020), Frick, Iijima, and Ishii (2019), Gagnon-Bartsch (2016), and Mailath and Samuelson (2019).

tion $a \in A$ induces an objective probability distribution $p^*_a \in \Delta(Y) \subset \mathbb{R}^{|Y|}$ over the finite set of possible outcomes Y.[8] Second, the action, paired with the realized outcome, determines the flow payoff of the agent via the utility function $u : A \times Y \to \mathbb{R}$.

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 $P = \bigtimes_{a \in A} \Delta(Y) \subset \mathbb{R}^{|Y| \times |A|}$ 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, and denote by $B_\varepsilon(p)$ the ball of radius ε around $p \in P$.[9]

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 topology of weak convergence of measures.

Definition 1. The *conceivable outcome distributions* are the elements of $\Theta = \operatorname{supp} \mu_0$. The agent is *correctly specified* if $p^* \in \Theta$, i.e. the objective distribution is conceivable.

Throughout the paper, we will maintain the following assumption:

Assumption 1 (Regularity)**.**

(i) For all $p \in \Theta$ and $a \in A$, $p_a(y) > 0$ if and only if $p^*_a(y) > 0$.

(ii) The prior μ_0 has *subexponential decay*: there is $\Psi : \mathbb{R}_+ \to \mathbb{R}$ such that for every $p \in \Theta$ and $\varepsilon > 0$ we have $\mu_0(B_\varepsilon(p)) \geqslant \Psi(\varepsilon)$ with $\lim \Psi(K/n) \exp(n) = \infty$ for all $K > 0$.

Assumption 1(i) requires that the set of outcomes that the agents thinks are possible coincides with the set of outcomes that objectively have positive probability. This assumption guarantees that Bayes rule is always well defined.[10] Assumption 1(ii) extends Diaconis and Freedman (1990)'s notion of ϕ-positivity to the misspecified case, and adds the requirement that the bounding Ψ vanishes at a subexponential rate around 0. It is always satisfied by priors with a density that is bounded away from 0 on their support, and by priors with finite support.[11]

[8] We denote objective distributions with a superscript $*$.

[9] For every finite dimensional vector v, we let $||v|| = \max_i v_i$ denote the supremum norm.

[10] Assumption 1(i) is satisfied in most applications but it is stronger than necessary. We explain in Online Appendix B.1 how our results extend to weaker assumptions on the support of the agent's prior beliefs.

[11] Dirichlet priors also satisfy Assumption 1(ii), even though they do vanish at the edge of their support. Fudenberg, He, and Imhof (2017) shows by example that even correctly specified Bayesian updating can behave oddly when the prior vanishes exponentially quickly.

Our specification allows the agent's subjective uncertainty to be correlated across actions. For example, in subjectively exogenous problems, 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 Subjective Beliefs We assume throughout that the agent updates their beliefs using Bayes rule. Denote by $\mu_t(\cdot \mid (a^t, y^t))$ the subjective belief the agent obtains using Bayes rule after action sequence $a^t = (a_s)_{s=1}^t$ and outcome sequence $y^t = (y_s)_{s=1}^t$,

$$\mu_t(C \mid (a^t, y^t)) = \frac{\int_{p \in C} \prod_{\tau=1}^t p_{a_\tau}(y_\tau) d\mu_0(p)}{\int_{p \in P} \prod_{\tau=1}^t p_{a_\tau}(y_\tau) d\mu_0(p)} . \tag{Bayes Rule}$$

Since the agent's prior has support Θ, their posterior belief does as well. We sometimes suppress the dependence of the posterior belief on the realized sequence and just write μ_t.

Behavior of the Agent A (pure) policy $\pi : \bigcup_{t=0}^\infty A^t \times Y^t \to A$ specifies an action for every history. 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 let $a_{t+1} = \pi(a^t, y^t)$ denote the action taken in period t. The objective action-contingent probability distribution p^* and a policy π induce a probability measure $\mathbb{P}_\pi$ on $(a_\tau, y_\tau)_{\tau=1}^\infty$.[12] Standard results guarantee that there is an optimal policy π that depends on the history only through the agent's beliefs; we restrict attention to policies that satisfy this restriction.

Given a belief $\nu \in \Delta(\Theta)$ we denote by ν_a the belief over outcome distributions associated with action a, i.e. $\nu_a(C) = \int \mathbf{1}_{p_a \in C} d\nu(p)$ for all Borel sets $C \subseteq \Delta(Y)$. 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. $A^m(\nu)$ denotes the set of myopically optimal actions given belief ν, i.e.,

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

2.2 Forms of Misspecification

Our model encompasses many sorts of misspecified learning, including the following:

[12] We spell out the details of this measure at the start of the Appendix.

Subjectively Exogenous Problems We say that there are subjectively exogenous outcomes when the agent believes that the realized outcome is not affected by the chosen action. More formally:

Definition 2. Outcomes are *subjectively exogenous* if for every $a, a' \in A$ and every $p \in \Theta$, we have $p_a = p_{a'}$.

Note that the agent can believe in exogenous outcomes independent of whether or not the action really does influence the distribution; if the action does influence the outcome and the agent ignores this we say the agent exhibits causation neglect. An agent who thinks the outcome distribution is exogenous updates their beliefs as if they faced an i.i.d. environment. We will establish that the beliefs in this setting concentrate on the conceivable outcome distributions closest to the empirical average. We use this result to show that if a is a uniformly strict Berk-Nash equilibrium, it is positively attractive.

Subjective Bandit Problems The other extreme case encompassed by our setup is where the agent thinks that they face a bandit problem, i.e. they believe that the distributions over outcomes induced by different actions are independent. This corresponds to the case where the agent's prior μ_0 is a product measure.

Definition 3 (Bandit Problem)**.** We say that an agent faces a *subjective bandit problem* if $\mu_0 = \bigtimes_{a \in A} \mu_{0,a} \in (\Delta(\Delta(Y)))^A$.

We show that uniformly strict Berk-Nash equilibria are positively attractive in this setting as well, provided that the agent is sufficiently patient.

One Dimensional Problems In one-dimensional problems, the agent's uncertainty is summarized by a parameter $\gamma \in \mathbb{R}$. The parameter determines the distribution over outcomes through a function ϕ which maps parameters to action-dependent outcome distributions. Formally, the support of the agent's prior μ_0 is contained in the image of this function ϕ.

Definition 4 (One-Dimensional Problems)**.** The problem is *one-dimensional* if there exists $\Gamma \subseteq \mathbb{R}$ and a function $\phi : \Gamma \to P$ such that $\Theta \subseteq \{\phi(\gamma) \colon \gamma \in \Gamma\}$. A one-dimensional problem is *supermodular* if A can be ordered such that $(\gamma, a) \mapsto \mathbb{E}_{\phi(\gamma)_a}[u(a, y)]$ is supermodular.

EPY provides a sufficient condition for actions to converge in one-dimensional problems that are supermodular. Heidhues, Kőszegi, and Strack (2018b) shows that a unique Berk-Nash equilibrium is globally attracting in supermodular problems where the outcomes are real numbers and ϕ is an additive shift. Our Example 9 shows that their result does not hold in our more general setting: a unique (and uniformly strict) Berk-Nash equilibrium may not be positively attractive. Under a stronger version of supermodularity, our positive attractiveness results do extend to extremal uniformly strict Berk-Nash equilibria.

Finite Support Another common assumption is that the support of the prior is finite. With a finite-support prior, if behavior converges to an action a, a is a best reply to all outcome distributions that minimize the Kullback-Leibler divergence from p^*_a, so it is a *uniform* Berk-Nash equilibrium. However, Example 6 shows that non uniform Berk-Nash equilibria can be limit points when the support of the prior is infinite if Assumption 1(ii) is not satisfied.

Signals Here we suppose that each period the agent observes a signal $s \in S$ 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 the 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 μ over P, where $p_{s,a}(y)$ denotes the probability under $p \in P$ of outcome y after observing signal s playing action a. Extending the model to signals lets us incorporate the stochastic payoff perturbations assumed in EP. It also lets us model cases where the agent mistakenly thinks that some information they observe is uninformative.

3 Limit Points and Berk-Nash Equilibria

We are interested in when the agent's actions converge, and their possible limit points. 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 the agent doesn't have an action choice, and conversely actions can oscillate with fixed beliefs if the agent is indifferent.[13]

We say that the action process *converges to action a* if there exists a time period $T \in \mathbb{N}$ such that $a_t = a$ for all time periods $t > T$. Action a is a *limit action* if the action process converges to a with positive probability under some optimal policy π.[14] Note that there may

[13] The fact that beliefs can oscillate under a fixed action is the driving force behind the uniformity requirement in several of our results, such as Theorem 1.

[14] Formally, there exists a measurable set $C \subseteq A^\infty \times Y^\infty$ with $\mathbb{P}_\pi[C] > 0$ such that a_t converges to a in C.

be several optimal policies for a given prior; 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 for which a is myopically optimal, and which assigns 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

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

Note that $-H(q, q')$ is the expected log likelihood of an outcome under subjective distribution q' when the true distribution is q, so q' with smaller $H(q, q')$ better explain q. The *Kullback-Leibler* (KL) divergence between p^*_a and p_a is given by $H(p^*_a, p_a) - H(p^*_a, p^*_a)$, so any p_a that minimizes $H\left(p^*_a, p\right)$ also minimizes the KL divergence between p^*_a and p_a.

Recall p_a denotes the outcome distribution that p assigns to action a. For each a, let

$$\hat{\Theta}(a) = \operatorname*{argmin}_{p \in \Theta} H\left(p^*_a, p_a\right) \subseteq \Theta \tag{1}$$

denote the set of conceivable action-contingent outcome distributions that minimize the KL divergence relative to the true distribution p^*_a given that the agent plays a. Note that the elements of $\hat{\Theta}(a)$ specify an outcome distribution for each action $a' \in A$, even though $\hat{\Theta}(a)$ only depends on the distributions corresponding to a. We call $\hat{\Theta}(a)$ the set of *KL minimizers* for action a.[15]

Berk (1966) established that the agent's beliefs concentrate on $\hat{\Theta}(a)$ if they always play a. This motivates Esponda and Pouzo (2016)'s notion of a Berk-Nash equilibrium. We introduce variations of this concept to capture different senses in which an action is or is not a long-run outcome of the agent's learning process.

Definition 5. Two action-contingent outcome distributions p and p' are *observationally equivalent under action a* if $p_a = p'_a$. We denote by $\mathcal{E}_a(p) \subseteq \Theta$ the set of action-contingent outcome distributions in Θ that are observationally equivalent to p under a.

Definition 6 (Equilibrium concepts)**.**

[15] Note that if $p^* \in \Theta$ then each minimizing p explains the observed outcome distribution perfectly, $p_a = p^*_a$. In particular this is true if μ_0 has full support.

(i) Action $a \in A$ is a *Berk-Nash equilibrium* (BN-E) if for some belief $\nu \in \Delta(\hat{\Theta}(a))$, a is myopically optimal given ν, i.e. $a \in A^m(\nu)$.

(ii) Action a is a *strict BN-E* if for some belief in $\nu \in \Delta(\hat{\Theta}(a))$, a is the unique myopically optimal action, i.e. $\{a\} = A^m(\nu)$.

(iii) Action a is a *uniform BN-E* if for all KL minimizers $p \in \hat{\Theta}(a)$ there exists a belief $\nu \in \Delta\left(\mathcal{E}_a(p)\right)$ such that $a \in A^m(\nu)$.

(iv) Action a is a *uniformly strict BN-E* if for every belief $\nu \in \Delta(\hat{\Theta}(a))$, a is the unique myopically optimal action, i.e., $\{a\} = A^m(\nu)$.

Uniformity requires that for each class of observationally equivalent KL minimizers for action a, there is a belief concentrated on that class for which a is the myopically optimal choice.[16] The difference between BN-E and uniform BN-E disappears in the correctly specified case, where both concepts coincide with self-confirming equilibrium. In settings where the KL minimizer is unique, the uniformity requirement has no bite. However, in frameworks with additional structure, such as symmetry or parametric restrictions, multiple KL minimizers can arise naturally. For example, suppose that agent's payoff depends on the color y of a ball drawn from an urn, and the agent's action is to bet on the color of the drawn ball. The agent correctly believes their action has no impact on the distribution of outcomes. The urn has 6 balls: 4 of them white, 1 red, 1 blue. Here there is a finite number of possible outcome distributions corresponding to the possible urn composition. If the agent wrongly believes that at most half of the balls share the same color, i.e., $p(y) \leqslant 1/2$ for $y \in \{$white, red, blue$\}$, the two KL minimizers are (3 white, 2 blue, 1 red) and (3 white, 1 blue, 2 red).

The following result motivates our definition of uniform BN-E. It holds regardless of the agent's discount factor, and for all optimal strategies. The same is true for all subsequent results except those where the dependence on the discount factor is made explicit.

Theorem 1. *Every limit action is a uniform BN-E.*

One implication of Theorem 1 is that limit actions must be BN-E. In outline, this follows from the fact that if actions converge to an action then eventually the agent always plays that action, 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.

[16] Notice that when $\mathcal{E}_a(p)$ is not a singleton for some $p \in \hat{\Theta}(a)$, uniformity does not require that the equilibrium action is a best reply to every KL minimizer in $\hat{\Theta}(a)$. The only other equilibrium refinement we know of that, like uniform BN-E, tests for optimality against all beliefs in a non-singleton set is Fudenberg and He (2020), which studies non-equilibrium learning in a steady-state model where the agents are correctly specified Bayesians. They do not study the dynamics away from the steady state.

More strongly, Theorem 1 shows that a limit action must be a *uniform* BN-E. When a is not a uniform BN-E, there is an equivalence class of KL minimizers such that a is not a myopic best reply when beliefs concentrate on that class. The proof of Theorem 1 works by contradiction: Consider an action a which is not a uniform BN-E. If play converges to a with positive probability there must exist a history after which it is optimal in every future period to play a. We thus study the agent's belief process under the assumption that a is played in every period. As we prove in Proposition 1 in the Appendix, the agent's beliefs concentrate around the set of Kullback-Leibler minimizers relative to the realized outcome frequency exponentially fast. This result allows us to determine the agent's long run actions from the long-run frequency of outcomes. If a is not a uniform BN-E, there is a KL minimizer p' under which action a is not optimal. Moreover, the number of times each outcome is realized is a random walk, and by the Central Limit Theorem the outcome frequency converges to objective outcome frequency p^*_a at rate $1/\sqrt{t}$. This implies that the probability with which the outcome frequency will be in a ball of radius $1/\sqrt{t}$ centered around $p^*_a\,(1 - 1/\sqrt{t}) + p'_a\,(1/\sqrt{t})$ in a given period t converges to a constant. These balls are chosen in the direction of the outcome frequency p'_a such that the action a is not optimal for large enough t when the agent's belief is in these balls. We then apply the Kochen-Stone Lemma which implies that the probability that the agent's outcome frequency will be in such a ball infinitely often is non-negative and the Hewitt-Savage zero-one law implies that it must equal one. Thus with probability one, the outcome frequency will eventually be such that the agent takes an action different from a. Thus, a can not be a limit action if it is not a uniform BN-E.

The same technique can be applied to obtain a starker result in subjective bandit problems. There Corollary 2 shows that if an action performs poorly under some KL minimizer, the agent will stop to play it in finite time with probability 1, even if the action is objectively optimal and the agent is very patient. Example 6 in the Online Appendix shows that Theorem 1 can fail without Assumption 1(ii). Here the agent's prior has countable support and assigns vanishingly low probability to distributions that are close to one of the KL minimizers. However, Assumption 1(ii) does not ensure that a uniform BN-E exists, as shown in the following example. As a consequence, actions need not converge.

Example 1 (Non-existence of Uniform BN-E)**.** *A monopolist is uncertain about the demand for their product. Every period it posts a price* $a \in \{3, 4, 5, 6, 7\}$*, and then a randomly selected consumer observes the price and decides whether to buy* $(y = 1)$ *or not buy the good* $(y = 0)$*. The monopolist's maximizes revenue* $u(a, y) = ay$*, and the true distribution of customer* *values is uniform on* $[3,7]$*. The monopolist overestimates the variance of consumer values, and believes that they are either uniformly distributed on* $[0,8]$ *or on* $[2,10]$*. As we show in the Online Appendix, the unique BN-E is nonuniform and strict, with price* 5*. Both distributions are KL-minimizing for this price, but price* 5 *is myopically optimal only if the valuations are uniformly distributed on the high range* $[2,10]$*. Theorem 1 implies that the monopolist's actions do not converge, even though there is a unique and strict BN-E. This is because when* $a = 5$*, the monopolist eventually sees a sequence of outcomes where few consumers buy, becomes very confident in the low range of valuations* $[0,8]$*, and switches to a lower price.*

Theorem 1 implies the non-convergence theorem of Nyarko (1991) as a corollary since also in that setting there is no uniform BN-E. Moreover, in the case of myopic agents, Corollary 3 in the Appendix combines the result with Theorem 2 of Esponda, Pouzo, and Yamamoto (2019) to show the empirical action frequencies cannot converge to some non-uniform BN-E.

4 Sufficient Conditions for Long-Run Persistence

Theorem 1 shows that play can only converge to a given action a if that action is a uniform BN-E. This section gives sufficient conditions for a to be a long-run outcome in two different senses, namely stability and attractiveness.

4.1 Stability

We say that action a is stable if play converges to a with high probability starting from every belief in a neighborhood of a KL minimizer for a. For $\nu \in \Delta(\Theta)$, let $B_\varepsilon(\nu) = \{\nu' \in \Delta(\Theta) | d(\nu', \nu) \leqslant \varepsilon\}$ be the set of beliefs over conceivable distributions that are within ε of ν. Define the set $\hat{\Theta}^\varepsilon(a)$ as all outcome distributions whose marginal distribution with respect to action a is at most ε away from a KL minimizer,

$$\hat{\Theta}^\varepsilon(a) = \{p \in \Theta \colon \text{ there exists } p' \in \hat{\Theta}(a) \text{ with } ||p'_a - p_a|| \leqslant \varepsilon\}. \tag{2}$$

Definition 7 (Stability)**.**

(i) An action a is *stable* if for every $\kappa \in (0,1)$, there is an $\varepsilon > 0$ and a belief $\nu \in \Delta(\Theta)$ such that for all initial beliefs in $B_\varepsilon(\nu)$, the action prescribed by some optimal policy converges to a with probability larger than $1 - \kappa$.

(ii) An action a is *uniformly stable* if for every $\kappa \in (0,1)$, there is an $\varepsilon > 0$ such that for all prior beliefs $\nu \in \Delta(\Theta)$ such that $\nu(\hat{\Theta}^{\varepsilon}(a)) > 1 - \varepsilon$, the action prescribed by any optimal policy converges to $a \in A$ with probability greater than $1 - \kappa$.

Theorem 1 shows that stable actions must be uniform BN-E. The next theorem shows that an action is a uniformly strict BN-E if and only if it is uniformly stable.

Theorem 2. *An action is uniformly stable if and only if it is a uniformly strict BN-E.*

Theorem 2 differs from past work by providing the first if and only if characterization of the stability of actions under misspecified learning with non-binary priors, and by allowing the agent to be non-myopic and thus perceive an information value from experimentation.[17] Its proof has two parts, corresponding to the two directions of the if and only if statement. To show that every uniformly strict BN-E is uniformly stable, we first show that if beliefs assign sufficiently high probability to a neighborhood of the KL minimizers, the only optimal action is the uniformly strict BN-E a. That such a neighborhood exists for a myopic policy follows from the definition of uniformly strict BN-E. 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 BN-E this value is strictly lower than that of a. Then we combine an observation from FII with a generalization of the arguments in Fudenberg and Levine (1992) and the Dubins' upcrossing inequality to guarantee that if the probability initially assigned to the neighborhood is sufficiently high, it is unlikely to drop below the threshold that makes action a suboptimal.

The proof of the converse direction is much simpler: If a is not a uniformly strict BN-E, there is a distribution p in $\hat{\Theta}(a)$ that makes some other action b the best response, and if we set ν to be a point mass on p the agent always plays b.

Theorem 2 is in contrast to the non-convergence in the monopoly pricing example of Heidhues, Kőszegi, and Strack (2018a), where there is a continuum of actions, and actions that are sufficiently near the strict best response are best responses to nearby beliefs. As we explain in Section 6, it is not clear what the right definition of uniform stability is for that setting.

[17] Bohren and Hauser (2020) and Fudenberg, Romanyuk, and Strack (2017) characterize stability when the agent has a binary prior. FII's Theorem 1 gives a sufficient condition for stability when the agent's prior has finite support. The statement of the theorem is for their general model, which takes the evolution of the belief process as a primitive, and does not describe the agent's actions, discount factor, or optimization. The paper's three applications all assume myopic choice.

Example 1 shows that Theorem 2 does not extend to strict BN-E that are not uniformly strict. The next example shows that in Theorem 2 we cannot replace uniformly stable with stable.

Example 2 (A stable BN-E that is not uniformly strict)**.** *Suppose there are 2 actions, a and b, that induce the same distribution on $Y = \{0, 1\}$ and such that $u(a, \cdot) = u(b, \cdot)$. The agent has an arbitrary belief supported on $\{p : p_a = p_b\}$, i.e., they know the actions induce the same distribution. Here, since the agent is always indifferent, even though action a is not a uniformly strict BN-E, it is stable under the (optimal) policy that prescribes to always play a.*

In general there is a gap between uniformly strict BN-E and stability, but in sufficiently rich problems, this gap is absent.

Definition 8. A problem is *rich* if for every action a, minimizer $p \in \hat{\Theta}(a)$ and $\varepsilon > 0$ there exists a $p' \in \Theta \backslash \hat{\Theta}(a)$ with $||p - p'|| \leqslant \varepsilon$ such that

$$\mathbb{E}_{p_a}\left[u(a, y)\right] - \max_{b \in A \backslash \{a\}} \mathbb{E}_{p_b}\left[u(b, y)\right] > \mathbb{E}_{p'_a}\left[u(a, y)\right] - \max_{b \in A \backslash \{a\}} \mathbb{E}_{p'_b}\left[u(b, y)\right].$$

In words, a problem is rich if for every KL minimizer for every action a, the agent's prior includes a nearby distribution under which a performs relatively less well.[18] This rules out the previous example and also rules out finite-support priors.

Theorem 3. *If a problem is rich, the following are equivalent:*

(i) a $\in A$ *is a uniformly strict BN-E.*

(ii) a $\in A$ *is stable.*

Richness guarantees that if a is not a uniformly strict equilibrium, there is a KL minimizer for action a that can be approximated with a sequence of outcome distributions $(p^n)_{n \in \mathbb{N}}$ under which action a is strictly suboptimal. To prove this theorem, for every ν we build a sequence of beliefs $(\nu^n)_{n \in \mathbb{N}}$ that have have p^n has the unique KL minimizer for action a, and combine this with Theorem 1 to show that the probability that the actions converge to a starting from ν_n is 0. To summarize our stability results,

Uniformly Strict BN-E = Uniformly Stable $\subseteq$ Stable $\subseteq$ Uniform BN-E,

where the first inclusion is an equality if the problem is rich.

[18] Note that "relatively less well" allows the action to be a best response to all distributions near p.

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* initial beliefs. Another natural notion of a being a long-run outcome is that for *every* initial belief with support Θ there is strictly positive probability that the agent's action converges to a.

Definition 9 (Positively attractive)**.** The action $a \in A$ is *positively attractive* if for every optimal policy π and every initial belief ν with $\operatorname{supp} \nu = \Theta$,

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

Below we give sufficient conditions for uniformly strict BN-E to be positively attractive. Benaïm and Hirsch (1999) obtains a similar conclusion for the linearly stable Nash equilibria of stochastic fictitious play.[19] These arguments rely on Proposition 1 in the Appendix, which shows that beliefs about the outcome distribution concentrate around the distributions that best fit the empirical frequency of outcomes. Importantly, our result applies pathwise and does not require that either actions or empirical frequencies converge.

Our results on positive attractiveness cover three different cases: subjectively exogenous outcomes, subjective bandit problems, and strongly supermodular problems. In the first two cases we are able to identify a particular empirical distribution that is sufficient for analyzing convergence. With subjectively exogenous outcomes, the agent only tracks a single empirical distribution. In subjective bandit problems, the agent does consider multiple empirical distributions, but it is sufficient to study the distribution corresponding to the action in question. In supermodular problems, we instead show that certain outcome realizations can lead the agent to lock on to the highest or lowest action.

4.2.1 Subjectively Exogenous Problems

In subjectively exogenous problems, the agent believes that the distribution over outcomes is the same for all actions. This is a fairly stark assumption; more typically the agent might believe that their action influences some dimensions of the outcome but not others. We present the case where the agent believes the action has no effect at all because the extension to "partially exogenous" outcomes does not bring any additional insight.

[19] The Bayesian foundation of fictitious play (Fudenberg and Kreps, 1993) assumes that the players believe that the environment is stationary. Away from a steady state the players are misspecified, but when the system converges to a steady state the stationarity assumption is asymptotically correct. In our setting, "substantial" misspecification can persist even when behavior converges.

Theorem 4. *Suppose outcomes are subjectively exogenous. If a is a uniformly strict BN-E such that p^*_{a} is absolutely continuous with respect to $p^*_{\mathrm{a}'}$ for all* $\mathrm{a}' \in A$*, then it is positively attractive.*

Example 9 in the Online Appendix shows that outside of subjectively exogenous problems, uniformly strict equilibria need not be positively attractive, even when they are the unique BN-E. To prove the result, we first use Proposition 1 to show that beliefs concentrate around the distributions that minimize the KL divergence from the empirical frequency on every path of outcome realizations. We then 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 BN-E, if beliefs concentrate around the minimizers, a becomes the unique best reply. While using a, the relative probability the agent assigns to distributions in $\hat{\Theta}(a)$ 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 $\hat{\Theta}(a)$ to show that, with positive probability, the agent will stick to action a forever.

Corollary 1. *Suppose that outcomes are subjectively exogenous, and that the true outcome distribution p^* has full support. Then every uniformly strict BN-E is positively attractive.*

Proposition 4 in EPY shows that for every uniformly strict BN-E a, there exists at least one prior with support equal to Θ under which the policy converges to a with positive probability. FII provides sufficient conditions for the system to converge with probability 1 to a specific BN-E from any initial belief. Our Theorem 4 concludes that every uniformly strict BN-E has positive probability of being the limit behavior starting from *every* initial prior without imposing conditions that imply global convergence to a specific outcome.

Example 3 (Stackelberg game perceived as Cournot)**.** *The agent is a seller who every period faces a competing seller randomly drawn from a large population. The agent first chooses whether to produce low output, $a = q_L = 1$, or high output, $a = q_H = 2$. The competitor sets their quantity y at q_L or q_H after observing the agent's action: If the agent chooses low output the competitor produces high output with probability* $2/3$*, while if the agent chooses high output competitor produces an high quantity with probability* $1/3$.[20] *The agent believes that the competitor chooses output without observing the agent's action, and that they choose an high output with some unknown probability p: $\Theta = \{p \in \Delta(\{q_L, q_H\})^{\{q_L, q_H\}} : p_{q_H}(q_H) = p_{q_L}(q_H)\}$. The true distribution is $p^*_{q_H}(q_H) = 1/3 = p^*_{q_L}(q_L)$.*

[20] The randomness could arise from a distribution over production costs in the population of competitors.

The demand function of the consumers is linear, and the agent has no production cost; the utility function of the agent is $u(a, y) = a(4.5 - a - y)$. High output is objectively optimal for the agent, and this is also a uniformly strict BN-E. However, low output is also a uniformly strict BN-E, supported by the wrong belief that the observed high level of production of the competitor would be the same even if the agent increased output. By Theorem 4 both actions have a positive probability of arising as limit outcomes starting from every initial prior.

Without the assumption of subjectively exogenous outcomes, uniformly strict BN-E need not be positively attractive, even if one maintains the full support assumption.

Example 4 (A uniform BN-E that is not positively attractive)**.** *A central bank decides whether to keep a flexible exchange rate, $a = f$, or peg the currency to the dollar, $a = c$. The outcome has two binary components, $y = (y^e, y^s)$, where y^e says whether the economy is in a boom, and y^s whether there is a speculative attack on the currency. The bank only cares about the outcome. It likes booms and dislikes speculative attacks: $u(f, y) = y^e, u(c, y) = \frac{3}{2}y^e - y^s$. The bank correctly believes that whether there is a speculative attack is independent of the state of the economy. Furthermore, the bank knows that if they maintain a flexible exchange rate, the probability of a currency attack is* 0*, and believes that the probability of a currency attack under a fixed exchange rate is either* 20% *(the true value) or* 90%*. The bank correctly believes that pegging the currency to the dollar increases the probability of a boom by* $33.\bar{3}\%$ *over a baseline probability, which the bank believes is either* $33.\bar{3}\%$ *or* $66.\bar{6}\%$*. In truth the baseline is* 50%*, so the bank is misspecified.*[21]

Here pegging the currency to the dollar is a uniformly strict BN-E, but it is not positively attractive: For any discount factor, if the prior assigns sufficiently high probability to the states where a currency attack happens with probability 90% *if the currency is not pegged to the dollar, the bank starts out choosing a flexible exchange rate, and sticks with that action forever. To see why, note that when the currency is floating the bank does not update its beliefs about the likelihood of a currency attack under a pegged exchange rate.*

4.2.2 Subjective Bandit Problems

Recall that in a subjective bandit problem (Definition 3), the agent believes that the outcome distribution is independent across actions. An argument similar to that for subjectively exogenous problems shows that uniformly strict BN-E are positively attractive in subjective

[21] That is, the bank believes that the probabilities of a boom with or without peg are either $(100\%, 66.\bar{6}\%)$ or $(66.\bar{6}\%, 33.\bar{3}\%)$, respectively, while in truth they are $(83.\bar{3}\%, 50\%)$.

bandit problems if the agent is sufficiently patient. However, uniformly strict BN-E is very demanding concept in subjective bandit problems, as the Kullback-Leibler divergence between the true and subjective outcome distributions induced by an action 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, in these problems we can replace the uniformity requirement with the requirement that the equilibrium is *weakly identified* introduced in Esponda and Pouzo (2016).

Definition 10. A BN-E a is *weakly identified* if for all $p, p' \in \hat{\Theta}(a)$ we have $p_a = p'_a$.

Weak identification guarantees that once behavior stabilizes on action a, there is no additional updating about the relative likelihood of the KL-minimizing outcome distributions. When the agent thinks the outcome distribution is exogenous, the equilibrium can only be weakly identified if 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_a that best matches p_a^*, without imposing any restrictions on what the agent believes about the consequences of other actions.

Theorem 5. *For every subjective bandit problem there is a $\bar{\beta} < 1$ such that if the discount factor $\beta \geqslant \bar{\beta}$, then every weakly identified strict BN-E is positively attractive.*

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

In subjective bandit problems, we can sharpen the conclusion of Theorem 1 for actions that perform poorly under one of the KL minimizers. We say that action a is *quasi-dominated* if there are $\hat{p} \in \hat{\Theta}(a)$ and $b \in A$ such that for all $p \in \Theta$ $\mathbb{E}_{\hat{p}_a}\left[u(a, y)\right] < \mathbb{E}_{p_b}\left[u(b, y)\right]$. That is, there is a KL minimizer $\hat{p}$ for action a such that the utility of a under $\hat{p}$ is lower than that of action b under any of the p in the support of the prior. Quasi-dominated actions are not uniform BN-E, so play cannot converge to them with positive probability.

In a subjective bandit problem even more is true; quasi-dominated actions can be played only a finite number of times.

Corollary 2. *In a subjective bandit problem, any quasi-dominated action is almost surely played only a finite number of times.*

In particular, in two-armed subjective bandit problems where one action is quasi-dominated, play converges to to the other one. Note that this result does not depend on the discount factor, and is true even if the quasi-dominated action is objectively optimal and the agent assigns positive probability to it being optimal. In contrast, the probability that a correctly specified agent locks on to an incorrect action goes to 0 as the discount factor goes to 1.

4.2.3 Strongly Supermodular problems

Definition 11. We say that the problem is *strongly supermodular* if we can strictly order the space of actions $(A, >)$, outcomes $(Y, >)$, and the set of conceivable distributions $(\Theta, >)$ so that:

(i) u is strictly supermodular in a and y;

(ii) if $p, p' \in \Theta$ and $p > p'$, then for all $a \in A$ and $y \in Y \backslash \bar{y}$, we have $p_a(\{y' : y' > y\}) > p'_a(\{y' : y' > y\})$, where $\bar{y}$ denotes the highest action.

Theorem 6. *In a strongly supermodular problem, if $p^*_{\underline{a}}$ (resp. $p^*_{\bar{a}}$) has full support, and the highest action $\bar{a}$ (resp. the lowest action $\underline{a}$) is a uniform and strict BN-E, then $\bar{a}$ (resp. $\underline{a}$) is positively attractive.*

Strong supermodularity implies that the agent will use action $\bar{a}$ if they observe the highest y's sufficiently often. Moreover, the antisymmetric ordering of the elements of Θ guarantees that every uniform and strict BN-E is uniformly strict, and so Theorem 2 guarantees that there is positive probability that once the agent plays $\bar{a}$ they will stick to it forever.

Example 5 (Under-investment trap)**.** *Each period the agent decides how much effort $a \in \{0, 1, 2\}$ to exert on a task. The effort can be either successful, $y = 1$,or unsuccessful, $y = 0$. Higher effort makes success more likely: $p^*_2(1) = 9/10 > p^*_1(1) = 1/2 > p^*_0(1) = 1/10$. Moreover, higher effort also increases the benefit of a success: $u(a, y) = ay - a/2$. Thus the objectively optimal action is to exert high effort, $a = 2$.*

The agent mistakenly believes that the probability of success depends on their effort and their intrinsic skill ψ, and Θ is consists of all p such that $p_2(1) = 2/3+\psi > p_1(1) = 1/2+\psi > p_0(1) = 1/3 + \psi$ for some $\psi \in [-1/4, 1/4]$.

Here there are two BN-E: $a = 0$ and $a = 2$. In the bad equilibrium $a = 0$, the KL-minimizing outcome distribution corresponds to the lowest possible skill level $\psi = -1/4$, which leads the agent to exert the low effort. Since both BN-E are uniformly strict and the problem is strongly supermodular, Theorem 6 implies that both the Nash equilibrium and the bad equilibrium with low effort are positively attractive.

5 Signals

Suppose each period before taking an action the agent observes a signal s from a compact set S, equipped with its Borel sigma algebra. Thus the analog of an action in the previous sections is now a *strategy*, i.e. a measurable 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 signals may also be useful for predicting the outcome distributions, so now $p_{a,s} \in \Delta(Y)$ depends both on this period's action and on the signal observed at the start of the period. A policy $\pi(a^t, y^t, s^{t+1})$ specifies the action in each period t as a function of past actions, outcomes and signals.

To complete the model we also need to specify the objective distribution of signals. We focus on the case where the distribution of s is fixed (iid) with distribution ζ that is known to the agent, as in Esponda and Pouzo (2016).[22]

Subjective Beliefs 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 $P = \Delta(Y)^{A\times S}$ be the space of all signal and action dependent outcome distributions. The agent's uncertainty is captured by a prior belief $\mu_0 \in \Delta\big(P\big)$, again with $\Theta = \operatorname{supp} \mu_0$.

Assumption 1′.

(i) For all $p \in \Theta$, $a \in A$, and $s \in S$, $p_{a,s}(y) > 0$ if and only if $p^*_{a,s}(y) > 0$.

(ii) The prior μ_0 has *subexponential decay*: there is $\Psi : \mathbb{R}_+ \to \mathbb{R}$ such that for every $p \in \Theta$ and $\varepsilon > 0$ we have $\mu_0(B_\varepsilon(p)) \geqslant \Psi(\varepsilon)$ with $\lim \Psi(K/n)\exp(n) = \infty$ for all $K > 0$.

Let $\mu_t(\cdot \mid (s^t, a^t, y^t)) \in \Delta(P)$ denote 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,

$$\mu_t(C \mid (s^t, a^t, y^t)) = \frac{\int_{p\in C} \prod_{\tau=1}^{t} p_{a_\tau, s_\tau}(y_\tau) d\mu_0(p)}{\int_{p\in P} \prod_{\tau=1}^{t} p_{a_\tau, s_\tau}(y_\tau) d\mu_0(p)}.$$

We say that two outcome distributions $p, p' \in \Theta$ are *observationally equivalent under the strategy* σ if $p_{\sigma(s),s}(y) = p'_{\sigma(s),s}(y)$ for all $y \in \operatorname{supp} p^*_{\sigma(s),s}$, and we let $\mathcal{E}_\sigma(p)$ denote the outcome distributions that are observationally equivalent to p under σ. To simplify the analysis, we make the following assumption, which is satisfied for example if the signals are payoff shocks, or if there is only finitely many signals.

[22] A continuum of signals allows payoff shocks that generate continuous best-response distributions.

Definition 12. The environment is *finite dimensional* if there is a partition $\Xi = \{\xi_1, ...\xi_N\}$ of S into a finite number of measurable sets such that the agent believes the same outcome distribution applies for all s in ξ_i: for all $p \in \Theta \cup \{p^*\}$, $a \in A$, and $s \in S$, $p_{a,s} = p_{a,s'}$ if $\xi(s) = \xi(s')$.

Under this assumption, we abuse the notation by letting p_{a,ξ_i} denote the outcome distribution prescribed by p after action a and an arbitrary signal in ξ_i. With this, the relevant set of "closest beliefs to the truth" is now

$$\hat{\Theta}(\sigma) = \underset{p \in \Theta}{\text{argmin}} \sum_{\xi_i \in \Xi} \zeta(\xi_i) H\left(p^*_{\sigma(s),\xi_i}, p_{\sigma(s),\xi_i}\right).$$

We use this modified definition of the minimizers to extend the definition of BN-E and uniformly strict BN-E to this more general setting. The extension to the case of finitely many signals is almost immediate. We allow for a continuum of payoff-relevant signals to be able to cover past work. This requires additional compactness arguments that do not provide additional insight about learning, so the proofs for all of the results of this section are in the Online Appendix.

Definition 13 (BN-E)**.**

(i) Strategy σ is a *BN-E* if there exists a belief $\nu \in \Delta(\hat{\Theta}(\sigma))$ such that σ is myopically optimal given ν.

(ii) Strategy σ is a *uniform BN-E* if for all $p \in \hat{\Theta}(\sigma)$ there exists a belief $\nu \in \Delta\left(\mathcal{E}_\sigma(p)\right)$ such that σ is myopically optimal given ν.

(iii) Strategy σ is a *uniformly strict BN-E* if σ is the unique myopic best reply to any belief in $\nu \in \Delta(\hat{\Theta}(\sigma))$.[23]

Theorem 1′. *Suppose the agent's beliefs are finite dimensional. Then if the strategy prescribed by the policy converges to σ with positive probability, then σ is a uniform BN-E.*

The proof of this result is very similar to the proof of Theorem 1. The main difference is that the relevant random walk is the empirical distribution over joint realizations of signals and outcomes.

Similarly, we can extend our result on the stability of uniformly strict BN-E. Specifically:

Theorem 2′. *Suppose σ is a uniformly strict BN-E. Then there is a belief $\nu \in \Delta(\Theta)$ such that for every $\kappa \in (0, 1)$ there exists an $\varepsilon' > 0$ such that starting from any prior belief in*

[23] Here uniqueness is up to a set of signals that have zero probability under ζ.

$B_{\varepsilon'}(\nu)$:

$$\mathbb{P}_{\pi}\left[\lim_{t\to\infty}\frac{1}{t+1}\sum_{r=0}^{t}\mathbf{1}_{\pi(a^r,y^r,s^{r+1})=\sigma(s_{r+1})}\geqslant 1-\kappa\right]>1-\kappa\,.$$

Example 10 in the Online Appendix illustrates the long-run biases that can be induced when the agent mistakenly thinks that signals are uninformative. There, a seller receives a signal about the current period's market, and decides whether to undertake an investment that may boost sales. The seller does not realize that when more consumers show up, a lower fraction of them buy; we show that this can lead to persistent underinvestment when market attendance is high.

When the agent thinks the signals are uninformative, their prior has support on distributions of y given a that are independent of s. Here the only reason they might influence the agent's choices is that they may directly enter their payoff function. The next result shows that all uniformly strict BN-E are positively attractive when signals are subjectively uninformative and the true data generating process has full support.

Theorem 4′. *If signals are finite, subjectively uninformative, outcomes are subjectively exogenous, and that the true outcome distribution p^* has full support, then any uniformly strict BN equilibrium σ is positively attractive.*

The proof of this result is similar to that of Theorem 4, because when signals are subjectively uninformative we can apply Proposition 1 to the *uncontingent* empirical distribution.

6 Concluding Remarks

Learning in Large Population Games The biases we consider are relevant in non-equilibrium models of learning about the prevailing distribution of strategies. Consider a finite I player game, and suppose there is a continuum of agents in each player role $i \in I$ who are matched every period to play the game, and observe the actions played in their matches but nothing else. In a steady state,[24] the problem faced by an agent in population i is equivalent to the one we 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's distribution of opponents' strategies. Causation neglect corresponds to the bias of an agent who thinks they are playing a simultaneous-move game, when in reality their

[24] These models do have steady states when there is a steady outflow of agents balanced by an inflow of new ones; see e.g. Proposition 3 in Fudenberg and He (2018).

opponents observe the agent's choice before moving. Subjective bandit problems arise when the agent has independent beliefs about the responses to different strategies. In games of incomplete information, the agent may have signal neglect, and incorrectly believe that the game has independent private values.

Our results help characterize the possible limit actions in these situations. Of course, extensive-form games may not have strict equilibria, so some of our results will not apply, but it may be possible to extend some of our 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.[25]

Markov Decision Problems If the agent's action influences the signal, then the true model is 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 the Markov structure, 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 future work.

Infinitely Many Actions When the agent has a finite number of possible actions or stage-game strategies, as we have assumed in this paper, an equivalent definition of uniformly strict BN-E is an action a that is the unique best response to every belief in a neighborhood of the KL minimizers for a. With infinitely many actions and continuous payoff functions, actions that are sufficiently near the strict best response incur arbitrarily small losses and are best responses to nearby beliefs. Here the two definitions of uniformly strict BN-E are not equivalent. Indeed, as shown by an example in Heidhues, Kőszegi, and Strack, 2018a, some BN-E that are uniformly strict BN in the sense of Definition 6 may not be positively attractive. However, we conjecture that the positive attractiveness result continues to hold under the alternative definition.

Summary and Discussion 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 these agents will develop and how they will behave. This paper

[25] Alternatively we could consider a model with one agent per player role and payoff perturbations, as in Fudenberg and Kreps (1993) and Esponda and Pouzo (2016).

provides sharp characterizations of what actions arise as the long-run outcomes of misspecified learning. We show that all uniformly strict BN-E are stable, and that under a mild condition only uniform BN-E can be stable. Moreover we show that play can only converge to uniform BN-E. Our work thus suggests uniformity should be imposed as a refinement of BN-E. We then provide the first sufficient conditions for an action to be positively attractive under misspecified learning. Here we highlight the role played by the correlation that the agent perceives between the outcome distributions associated with different actions.

A Appendix

Section A.1 formally describes the space where our stochastic processes are defined, Section A.2 states some preliminary technical lemmas, Section A.3 proves that beliefs concentrate around the KL minimizers at and exponential rate, and Section A.4 contains the results of the main text for the models that do not have signals.

A.1 State Space

We work with the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The state space $\Omega = (Y^\infty)^A$ consists of infinite sequences of action dependent outcome realizations $(x_{a,1}, x_{a,2}, \ldots)_{a \in A}$, where $x_{a,k}$ 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^*. The outcome observed by the agent in period t after action a_t is $y_t = x_{a_t,k}$, where $k = |\{\tau \leqslant t \colon a_\tau = a_t\}|$ is the number of times the agent has taken action a_t up to and including period t.[26] The probability measure $\mathbb{P}_\pi$ over $(a_\tau, y_\tau)_{\tau=1}^\infty$ induced by the policy π is defined as follows: For every $t \in \mathbb{N}$ and cylinder $(a_\tau, y_\tau)_{\tau=1}^t$,

$$\mathbb{P}_\pi\left[(a_\tau, y_\tau)_{\tau=1}^t\right] = \begin{cases} 0, & \text{if there exists } t' \in \{1, ..., t\} : a_{t'} \neq \pi((a_\tau, y_\tau)_{\tau=1}^{t'-1}) \\ \prod_{\tau=1}^t p_{a_\tau}(y_\tau) & \text{otherwise.} \end{cases}$$

[26] Several different state spaces lead to the same law for the stochastic processes we are interested in. In particular, we could have specified the more natural 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 in period k. An argument similar to that of Lemma 5 of Fudenberg and He (2017) shows that this choice would not change our results.

A.2 Preliminary Lemmas and Definitions

Denote the set of conceivable outcome distributions for action a that best match p_a^* by

$$\hat{\Theta}_a(a) = \operatorname*{argmin}_{p_a : p \in \Theta} H\left(p_a^*, p_a\right) \subset \Delta(Y).$$

Lemma 1. *For every $a \in A$ and $\varepsilon > 0$, $\hat{\Theta}(a)$ defined in (1), $\hat{\Theta}_a(a)$, $\hat{\Theta}^{\varepsilon}(a)$ defined in (2), and $\Delta(\hat{\Theta}(a))$ are compact.*

The proof of Lemma 1 is routine and relegated to the Online Appendix.

For every $p \in P$ and every policy π let $\mathbb{E}_{p,\pi}[\cdot]$ denote the expectation operator over action and outcome sequences that is induced by policy π under outcome distribution p. We work with the agent's normalized value throughout, which is

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

The set of policy functions is $\Pi = A^{\bigcup_{t=0}^{\infty} A^t \times Y^t}$.

Lemma 2. *Π is compact in the product topology, and for all $\nu \in \Delta\left(\Theta\right)$, $V\left(\cdot, \nu\right)$ is continuous with respect to the product topology.*

Lemma 2 is a consequence of the more general Lemma 11 which covers cases where each period the agent observes a signal before choosing their action. This lemma is proved in the Online Appendix.

Next we 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 $\hat{\Theta}(a)$. Denote the set of beliefs over conceivable distributions that assign at least probability $1-\varepsilon$ to $\hat{\Theta}^{\varepsilon}(a)$ by

$$M_{\varepsilon,a} = \{\nu \in \Delta(\Theta) \colon \nu(\hat{\Theta}^{\varepsilon}(a)) \geqslant 1-\varepsilon\}.$$

The following lemma shows that if the agent's beliefs are sufficiently concentrated on the set of KL minimizers associated with a uniformly strict BN-E a, the agent will play a, even if the agent is not myopic.

Lemma 3. *If $a \in A$ is a uniformly strict BN-E, for every optimal policy π, there exists an $\hat{\varepsilon} > 0$ such that for all $\varepsilon < \hat{\varepsilon}$, $\nu \in M_{\varepsilon,a} \implies \pi\left(\nu\right) = a$.*

Proof. Let π^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 belief ν in $M_{\varepsilon,a}$

$$G(\varepsilon) = \min_{\tilde{\pi}:\tilde{\pi}(\nu)\neq a} \min_{\nu \in M_{\varepsilon,a}} \left(V\left(\pi^a, \nu\right) - V\left(\tilde{\pi}, \nu\right)\right).$$

Notice that by Lemma 2, the space of the policy functions endowed with the product topology is compact. Since the subset of policy functions that do not prescribe a at the initial history is closed, this subset is compact as well, and because $\beta \in [0, 1)$, the value function is continuous at infinity, and therefore $V\left(\pi^a, \nu\right) - V\left(\cdot, \nu\right)$ is a continuous function of the policy. Notice also that since $\mathbb{E}_{p,\pi}\left[\sum_{t=1}^{\infty}\left[\beta^{t-1}u(a_t, y_t)\right]\right]$ is continuous in p, $V\left(\pi^a, \cdot\right) - V\left(\tilde{\pi}, \cdot\right)$ is continuous in ν. Therefore, given that $\varepsilon \to M_{\varepsilon,a}$ is an upper hemicontinuous and compact valued correspondence, we can conclude by the Maximum Theorem that G is continuous in ε. We have that

$$\begin{aligned}
G(0) &= \min_{\tilde{\pi}:\tilde{\pi}(\nu)=a} \min_{\nu \in \Delta(\hat{\Theta}(a))} \left(V\left(\pi^a, \nu\right) - V\left(\tilde{\pi}, \nu\right)\right) \\
&\geqslant \min_{\nu \in \Delta(\hat{\Theta}(a))} \left[V\left(\pi^a, \nu\right) - \max_{\tilde{\pi}:\tilde{\pi}(\nu)\neq a} V\left(\tilde{\pi}, \nu\right)\right] \\
&\geqslant \min_{\nu \in \Delta(\hat{\Theta}(a))} \left[(1-\beta)\int_P \sum_{y \in Y} p_a(y)u(a,y) - \max_{a' \neq a} \sum_{y \in Y} p_{a'}(y)u(a',y)d\nu(p)\right] \\
&\quad + \beta \min_{\nu \in \Delta(\hat{\Theta}(a))} \left[\max_{\tilde{\pi}} \frac{\sum_{t=1}^{\infty} \beta^{t-1} \int_P \left[\sum_{y \in Y} p_a(y)u(a,y) - \mathbb{E}_{p,\tilde{\pi}}\left[u(a_t, y_t)\right]\right] d\nu(p)}{1-\beta}\right] > 0
\end{aligned}$$

where the weak inequality follows by the concavity of the minimum and the strict inequality by the fact the unifom strictness of a. Therefore, there is an $\hat{\varepsilon}$ such that if $\varepsilon \leqslant \hat{\varepsilon}$, $G\left(\varepsilon\right) > 0$. This implies that for any optimal policy π it must be such that $\nu \in M_{\varepsilon,a}$ implies that $\pi(\nu) = a$, which proves the lemma. ■

The next Lemma extends an argument of Fudenberg and Levine (1992) to take into account misspecification. It establishes that if the expectation of the l-th power of the likelihood ratio between two subjective outcome distributions is greater 1 then the l-th power of the likelihood ratio of the subjective probability assigned to small environments of these outcome distributions is a sub-martingale.

Lemma 4. *Let $p, p', p^* \in \Delta(Y)$, and $l \in (0,1)$ be such that*

$$\sum_{y \in Y} p^*(y) \left(\frac{p(y)}{p'(y)}\right)^l < 1. \tag{3}$$

Then there is $\varepsilon' > 0$ such that for all $\nu \in \Delta(\Delta(Y))$, if we let $\nu(C \mid y) = \frac{\int_{q \in C} q(y) d\nu(q)}{\int_{q \in \Delta(Y)} q(y) d\nu(q)}$, then

$$\sum_{y \in Y} p^*(y) \left[\left(\frac{\nu(B_{\varepsilon'}(p) \mid y)}{\nu(B_{\varepsilon'}(p') \mid y)}\right)^l\right] \leqslant \left(\frac{\nu(B_{\varepsilon'}(p))}{\nu(B_{\varepsilon'}(p'))}\right)^l.$$

Proof. The lemma is trivially true if $\nu(B_\varepsilon(p')) = 0$ for some ε. Therefore, without loss of generality, we can assume that $\nu(B_\varepsilon(p')) > 0$ for all ε. Let $\hat{\varepsilon}$ be such that $||q - p'|| \leqslant \hat{\varepsilon}$ implies that $q(y) = 0$ only if $p^(y) = 0$. Let $C_\varepsilon = \Delta(B_\varepsilon(p)) \times \Delta(B_\varepsilon(p'))$ and define $G : [0, \frac{\hat{\varepsilon}}{2}] \to \mathbb{R}$ by

$$G(\varepsilon) = \max_{(\bar{\nu}, \nu') \in C_\varepsilon} \sum_{y \in Y} p^*(y) \left(\frac{\int_{B_\varepsilon(p)} \bar{q}(y) d\bar{\nu}(\bar{q})}{\int_{B_\varepsilon(p')} q(y) d\nu'(q)}\right)^l.$$

By the Maximum Theorem, the compactness of $\Delta(B_\varepsilon(p'))$ and $\Delta(B_\varepsilon(p))$ and the fact that $G(0) < 1$ by equation (3), there is $\varepsilon' > 0$ such that for all $\nu' \in \Delta(B_{\varepsilon'}(p'))$, $\bar{\nu} \in \Delta(B_{\varepsilon'}(p))$

$$\sum_{y \in Y} p^*(y) \left(\frac{\int_{B_{\varepsilon'}(p)} \bar{q}(y) d\bar{\nu}(\bar{q})}{\int_{B_{\varepsilon'}(p')} q(y) d\nu'(q)}\right)^l \leqslant 1. \tag{4}$$

Then

$$\begin{aligned}
\sum_{y \in Y} p^*(y) \left(\frac{\nu(B_{\varepsilon'}(p) \mid y)}{\nu(B_{\varepsilon'}(p') \mid y)}\right)^l &= \sum_{y \in Y} p^*(y) \left(\frac{\int_{B_{\varepsilon'}(p)} \nu(B_{\varepsilon'}(p)) \bar{q}(y) d\frac{\nu(\bar{q})}{\nu(B_{\varepsilon'}(p))}}{\int_{B_{\varepsilon'}(p')} \nu(B_{\varepsilon'}(p')) q(y) d\frac{\nu(q)}{\nu(B_{\varepsilon'}(p'))}}\right)^l \\
&= \sum_{y \in Y} p^*(y) \left(\frac{\int_{B_{\varepsilon'}(p)} \bar{q}(y) d\frac{\nu(\bar{q})}{\nu(B_{\varepsilon'}(p))}}{\int_{B_{\varepsilon'}(p')} q(y) d\frac{\nu(q)}{\nu(B_{\varepsilon'}(p'))}}\right)^l \left(\frac{\nu(B_{\varepsilon'}(p))}{\nu(B_{\varepsilon'}(p'))}\right)^l \\
&\leqslant \left(\frac{\nu(B_{\varepsilon'}(p))}{\nu(B_{\varepsilon'}(p'))}\right)^l
\end{aligned}$$

where the inequality follows from equation (4). ■

Lemma 5. *For any action b and sequence of actions a^t let π^{b,a^t} be the policy that prescribes action a_t at period t and action b at all periods $\tau > t$. If, for every t, every sequence of actions*

a^t, and every optimal policy $\tilde{\pi}$ almost surely the belief process under the policy π^{b,a^t} arrives at a belief μ_τ, $\tau \geqslant t$ where $\tilde{\pi}$ does not prescribe b, i.e. $\mathbb{P}_{\pi^{b,a^t}}[b = \tilde{\pi}(\mu_\tau)$ for all $\tau \geqslant t] = 0$ then b is not a limit action.

Proof. Fix an arbitrary optimal policy $\tilde{\pi}$ and an history (a^t, y^t) with $\mathbb{P}_{\tilde{\pi}}(a^t, y^t) > 0$ such that $b = \tilde{\pi}(\mu_t(\cdot|(a^t, y^t)))$. Let $\tau = \min\{s > t \colon b \neq \tilde{\pi}(\mu_t))\}$ be the first time after (a^t, y^t) when $\tilde{\pi}$ does not prescribe b. Note that since $\tilde{\pi}(a^s, y^s) = b = \pi^{b,a^t}(a^s, y^s)$ for all $s \in [t, \tau - 1]$ the agent's belief until period τ is the same under π^{b,a^t} and $\tilde{\pi}$. As $\mathbb{P}_{\tilde{\pi}}(a^t, y^t) > 0$ implies $\mathbb{P}_{\pi^{b,a^t}}(a^t, y^t) > 0$, the probability that the agent takes the action b forever (i.e. $\tau = \infty$) after history (a^t, y^t) equals 0 by the assumption of the lemma. So, under every arbitrary optimal policy, after every history where b is played an action different from b is played with probability 1, and therefore b is not a long-run action. ■

The next lemma extends Lemma 3 of FII to show that there exists a uniform l such that all KL minimizers dominate all the distributions that are ε away from the minimizers in the sense that the expectation of the l-th power of the likelihood ratio is lower than 1.

Lemma 6. *Fix an action a and $\varepsilon > 0$. There exists an $\bar{l} > 0$ such that for all $l \leqslant \bar{l}$ for every KL minimizer $q \in \hat{\Theta}(a)$ and every outcome distribution $p' \notin \hat{\Theta}^\varepsilon(a)$*

$$f_l\,(q, p') := \sum_{y \in Y} p^*_a(y) \left(\frac{p'_a(y)}{q_a(y)}\right)^l < 1\,.$$

Proof. As noted by FII in their Lemma 3, (i) for each KL minimizer $q \in \hat{\Theta}(a)$ and every outcome distribution $p' \notin \hat{\Theta}(a)$ there exists an $l\,(q, p')$ such that $f_l(q, p') < 1$ for all $l \leqslant l\,(q, p')$ and (ii) for all $q, q' \in \Theta$, if $\hat{l} > l$ and $f_l(q, q') \geqslant 1$, then $f_{\hat{l}}(q, q') \geqslant 1$. We will now prove that there exists a uniform l that works for every $q \in \hat{\Theta}(a)$ and $p' \notin \hat{\Theta}^\varepsilon(a)$.

Suppose by way of contradiction that there was no $\bar{l} > 0$ such that for all $l \leqslant \bar{l}$, $f_l(q, p') < 1$ for all $q \in \hat{\Theta}(a)$ and $p' \notin \hat{\Theta}^\varepsilon(a)$. Then define a sequence (q_n, p'_n) such that $f_{\frac{1}{n}}(q_n, p'_n) \geqslant 1$. Sequential compactness of $\hat{\Theta}(a) \times \mathrm{cl}\{p \in \Delta(\Theta) \colon p_a \notin \hat{\Theta}^\varepsilon(a))\}$ guarantees that this sequence has an accumulation point (q, p') with $q \in \hat{\Theta}(a)$ and $p' \notin \hat{\Theta}(a)$.[27] However, for $n > \frac{1}{l(\bar{p}, p')}$, $f_{\frac{1}{n}}(q_n, p'_n) \geqslant 1$ implies $f_{l(q,p')}(q_n, p'_n) \geqslant 1$, and the continuity of $f_{l(q,p')}$ at (q, p') leads to a contradiction with $f_{l(q,p')}\,(q, p') < 1$. ■

[27] We denote the closure of a set by cl.

A.3 Exponential Concentration of Beliefs

We show here that repeated use of action a implies that the beliefs about the outcome distribution induced by a concentrate at an exponential rate around the distributions that best fit the empirical frequency of observed outcomes. Importantly, this result does not require that either actions or empirical frequencies converge. It will be important in what follows that these results apply pathwise, as they do in the correctly specified case studied by Diaconis and Freedman (1990), although unlike their result ours only applies for empirical distributions that are near the true distribution p^*. For brevity, we limit our analysis to this set of distribution, since this is enough for our results. In a separate note, Fudenberg, Lanzani, and Strack (2020), we provide a result that resembles more closely the original result in Diaconis and Freedman (1990).

For every $a \in A$, $\eta \in (0,1)$ and $q \in \Delta(Y)$, let $f_{\eta,q} = (1-\eta)p_a^* + \eta q$, $\eta_t = 2t^{-\frac{1}{2}}$, and $D = \min\left\{(p'_a(y)/p_a(y)) : p, p' \in \Theta, a \in A, y \in Y, p_a^*(y) > 0\right\}.$

Proposition 1. *Let $(a_i, y_i)_{i=1}^{\tau}$ be a history with positive probability, and suppose that only action a is played in periods $(\tau+1, ..., \tau+t)$. For every $\hat{q} \in \hat{\Theta}_a(a)$, there exist $I, \hat{K}, K' \in \mathbb{R}_{++}$ such that if the empirical outcome frequency $f_t = \frac{1}{t}\sum_{i=\tau+1}^{\tau+t} \mathbf{1}_{y_i=y}$ satisfies $||f_{\eta_t,\hat{q}} - f_t|| < ||\hat{q} - p_a^*||t^{-\frac{1}{2}}/K'$ then*

$$\frac{\mu_{\tau+t}\left(\{p \in \Theta \colon \forall y \in \operatorname{supp} p_a^*, |p_a(y) - \hat{q}(y)| < \varepsilon\}\right)}{1 - \mu_{\tau+t}\left(\{p \in \Theta \colon \forall y \in \operatorname{supp} p_a^*, |p_a(y) - \hat{q}(y)| < \varepsilon\}\right)} \geqslant D^{\tau}\Psi\left(\hat{K}\varepsilon^2 \frac{2}{It^{\frac{1}{2}}}\right) \exp\left(2\hat{K}t^{\frac{1}{2}}\varepsilon^2\right).$$

To establish Proposition 1 we first prove a sequence of auxiliary results. Given two outcome distributions $q, q' \in \Delta(Y)$, $\eta \in (0,1)$, and $\varepsilon > 0$, let

$$U_\varepsilon(q, q', \eta) = \{q'' \in \Delta(Y) : ||\eta q + (1-\eta)q' - q''|| \leqslant \varepsilon\}$$

denote the ball of radius ε around $\eta q + (1-\eta)q'$. The next result establishes a form of local Lipschitz continuity of the function $\min_{q' \in C} H(\cdot, q') - H(\cdot, q)$ for suitably chosen $q \in \Delta(Y)$ and compact $C \subseteq \Delta(Y)$.

Lemma 7. *Fix $q \in \Delta(Y)$ with $\operatorname{supp} q \subseteq \operatorname{supp} p_a^*$ and a compact set $C \subseteq \Delta(Y)$ that contains a $\tilde{q}$ with $\operatorname{supp} p_a^* \subseteq \operatorname{supp} \tilde{q}$. Then there exists a $K > 0$ such that for every $f' \in U_\varepsilon(q, p_a^*, \eta)$ with $\operatorname{supp} f' \subseteq \operatorname{supp} p_a^*$*

$$\left|\min_{q' \in C} H\left((1-\eta)p_a^* + \eta q, q'\right) - H\left((1-\eta)p_a^* + \eta q, q\right) - \min_{q' \in C} H\left(f', q'\right) + H\left(f', q\right)\right| \leqslant K\varepsilon.$$

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

$$Q_{\varepsilon,\chi}(\bar{q}) = \left\{ q' \in \Delta(Y) : \exists q'' \in \underset{q \in \operatorname{supp} \chi}{\operatorname{argmin}} H\left(\bar{q}, q\right), ||q' - q''||_\infty < \varepsilon \right\}$$

be the distributions that are within ε of a distribution q'' that minimizes the Kullback-Leibler divergence with the given $\bar{q}$, and let

$$g\left(p', \varepsilon\right) = \min_{p \notin Q_{\varepsilon,\chi_0}(p')} H\left(p', p\right) - \min_{p \in \operatorname{supp} \chi_0} H\left(p', p\right) > 0$$

be the minimal increase of the relative entropy from p' when it is minimized over Q_{ε,χ_0} instead of $\operatorname{supp} \chi_0$.

Lemma 8. *Let $\chi_0 \in \Delta(\Delta(Y))$ and suppose that for every $t \in \mathbb{N}$, $C \subseteq \Delta(Y)$, and sequence of outcomes $y^t \in Y^t$*

$$\chi_t(C \mid y^t) = \frac{\int_{q \in C} \prod_{\tau=1}^t q(y_\tau) d\chi_0(q)}{\int_{q \in \Delta(Y)} \prod_{\tau=1}^t q(y_\tau) d\chi_0(q)}.$$

Then for all $\varepsilon > 0$

$$p'(y) = \frac{\sum_{\tau=1}^t \mathbf{1}_{y_\tau = y}}{t} \implies \frac{\chi_t\left(Q_{\varepsilon,\chi_0}\left(p'\right) \mid y^t\right)}{1 - \chi_t\left(Q_{\varepsilon,\chi_0}\left(p'\right) \mid y^t\right)} \geqslant \chi_0\left(Q_{\frac{g(p',\varepsilon)}{2R(p',\varepsilon)},\chi_0}\left(p'\right)\right) e^{.5tg(p',\varepsilon)}$$

where

$$R(p', \varepsilon) = \sup_{q,q' \in Q_{\varepsilon,\chi_0}(p')} \frac{|H(p', q) - H(p', q')|}{||q - q'||}.$$

Proof. Let $p'(y) = \left(\sum_{\tau=1}^t \mathbf{1}_{y_\tau = y}\right)/t$, fix $\varepsilon > 0$ and for any $\bar{\varepsilon} > 0$, let $Q(\bar{\varepsilon}) = Q_{\bar{\varepsilon},\chi_0}\left(p'\right)$. By definition of $R(p', \varepsilon)$, $\min_{p \notin Q(\varepsilon)} H\left(p', p\right) - \max_{p \in Q\left(\frac{g(p',\varepsilon)}{2R(p',\varepsilon)}\right)} H\left(p', p\right) \geqslant .5g\left(p', \varepsilon\right)$. From the definition of χ_t we have that for all y^t where the empirical distribution is p',

$$\begin{aligned}
\frac{\chi_t\left(Q(\varepsilon) \mid y^t\right)}{1 - \chi_t\left(Q(\varepsilon) \mid y^t\right)} &= \frac{\int_{Q(\varepsilon)} \sum_{y \in Y} q(y)^{tp'(y)} \left(1 - q(y)\right)^{t(1-p'(y))} d\chi_0(q)}{\int_{\operatorname{supp} \chi_0 \setminus Q(\varepsilon)} \sum_{y \in Y} q(y)^{tp'(y)} \left(1 - q(y)\right)^{t(1-p'(y))} d\chi_0(q)} \\
&\geqslant \frac{\int_{Q\left(\frac{g(p',\varepsilon)}{2R(p',\varepsilon)}\right)} \exp(-tH\left(p', q\right)) d\chi_0(q)}{\exp(-t \min_{p \notin Q(\varepsilon)} H\left(p', p\right))} \\
&= \int_{Q\left(\frac{g(p',\varepsilon)}{2R(p',\varepsilon)}\right)} \exp(t \min_{p \notin Q(\varepsilon)} H\left(p', p\right) - tH\left(p', q\right)) d\chi_0(q) \\
&\geqslant \chi_0\left(Q\left(\frac{g(p', \varepsilon)}{2R(p', \varepsilon)}\right)\right) e^{.5tg(p',\varepsilon)},
\end{aligned}$$

where the first inequality follows from $g(p', \varepsilon)/2R(p', \varepsilon) \leqslant \varepsilon$. ■

Lemma 9. *For $\varepsilon > 0$ and $\eta \in (0,1)$, if $q = p_a, p \in \hat{\Theta}(a)$ then $g\left((1-\eta)p_a^* + \eta q, \varepsilon\right) \geqslant 2\eta\varepsilon^2$.*

Proof. H is linear in its first argument, so for $\eta \in (0,1)$, $\operatorname{argmin}_{p_a : p \in \Theta} H((1-\eta)p_a^* + \eta q, p_a) = \{q\}$. Then

$$\begin{aligned}
&g\left((1-\eta)p_a^* + \eta q, \varepsilon\right) \\
&\quad \geqslant \inf_{q' \in \Delta(Y)\setminus B_\varepsilon(q)} \sum_{y \in Y} \left[(1-\eta)p_a^*(y) + \eta q(y)\right] \log q'(y) - \sum_{y \in Y} \left[(1-\eta)p_a^*(y) + \eta q(y)\right] \log q(y) \\
&\quad \geqslant (1-\eta) \inf_{q' \in \Delta(Y)\setminus B_\varepsilon(q)} \sum_{y \in Y} p_a^*(y) \left[\log q'(y) - \log q(y)\right] \\
&\qquad\qquad\qquad\qquad + \eta \inf_{q' \in \Delta(Y)\setminus B_\varepsilon(q)} \sum_{y \in Y} q(y) \left[\log q'(y) - \log q(y)\right] \\
&\quad \geqslant 0 + \eta \inf_{q' \in \Delta(Y)\setminus B_\varepsilon(q)} \sum_{y \in Y} q(y) \left[\log q'(y) - \log q(y)\right] \geqslant 2\eta\varepsilon^2,
\end{aligned}$$

where the first inequality follows from the definition of g and the fact that the RHS minimizes over a larger set, the second inequality follows from concavity of the minimum, the third from the fact that q is a KL minimizer, and the fourth from Corollary 3.5 and Proposition 4.7 in Diaconis and Freedman (1990). ■

Remark 1. Observe that after every finite time t, the posterior μ_t almost surely satisfies the regularity assumption. That (i) is satisfied follows from the fact that $\operatorname{supp} \mu_t \subseteq \operatorname{supp} \mu_0$. For (ii), let $\Psi : \mathbb{R}_+ \to \mathbb{R}$ be the function whose existence is guaranteed by the regularity assumption (ii). Bayesian updating implies that for every $p \in \Theta$, $\varepsilon > 0$, $\mu_t(B_\varepsilon(p)) \geqslant \mu_0(B_\varepsilon(p))D^t \geqslant \Psi(\varepsilon)D^t$ a.s. Therefore, by defining $\Psi_t = D^t\Psi$ we have $\lim_{n\to\infty} \Psi_t(K/n)\exp(n) = \lim_{n\to\infty} \Psi(K/n)\exp(n)D^t = \infty$ for all $K > 0$, so (ii) is satisfied.▲

Proof of Proposition 1. Set $I = R(f_{\eta_t, \hat{q}}, \varepsilon)$. If $f_{\eta_t} = \frac{\sum_{i=\tau+1}^{\tau+t} \mathbf{1}_{y_i = y}}{t}$ there exists an $I \in \mathbb{R}_+$ such that

$$\begin{aligned}
&\frac{\mu_{\tau+t}\left(\{p \in \Theta \colon \forall y \in \operatorname{supp} p_a^*, |p_a(y) - \hat{q}(y)| < \varepsilon\}\right)}{1 - \mu_{\tau+t}\left(\{p \in \Theta \colon \forall y \in \operatorname{supp} p_a^*, |p_a(y) - \hat{q}(y)| < \varepsilon\}\right)} \\
\geqslant \quad & \mu_\tau\left(\{p \in \Theta \colon \forall y \in \operatorname{supp} p_a^*, |p_a(y) - \hat{q}(y)| < \frac{g(f_{\eta_t, \hat{q}}, \varepsilon)}{2I}\}\right) e^{.5tg\left(f_{\eta_t, \hat{q}}, \varepsilon\right)} \\
\geqslant \quad & \mu_\tau\left(\{p \in \Theta \colon \forall y \in \operatorname{supp} p_a^*, |p_a(y) - \hat{q}(y)| < \varepsilon^2 \frac{2}{It^{\frac{1}{2}}}\}\right) \exp\left(t\eta_t\varepsilon^2\right) \geqslant D^\tau \Psi\left(\varepsilon^2 \frac{2}{It^{\frac{1}{2}}}\right) \exp\left(2t^{\frac{1}{2}}\varepsilon^2\right),
\end{aligned}$$

where the first inequality follows from Lemma 8, the second from Lemma 9, and the third from Assumption 1(ii), and Remark 1.

By Lemma 7 there exists a $\hat{K}, K' > 0$ such that if $||f_{\eta_t,\hat{q}} - f_t|| < ||\hat{q} - p_a^*||t^{-\frac{1}{2}}/K'$ then

$$\frac{\mu_{\tau+t}\left(\{p \in \Theta \colon \forall y \in \operatorname{supp} p_a^*, |p_a(y) - \hat{q}(y)| < \varepsilon\}\right)}{1 - \mu_{\tau+t}\left(\{p \in \Theta \colon \forall y \in \operatorname{supp} p_a^*, |p_a(y) - \hat{q}(y)| < \varepsilon\}\right)} \geqslant D^\tau \Psi\left(\hat{K}\varepsilon^2 \frac{2}{It^{\frac{1}{2}}}\right) \exp\left(2\hat{K}t^{\frac{1}{2}}\varepsilon^2\right).$$

■

A.4 Proof of Results Stated in the Text

Proof of Theorem 1. We prove the statement by contraposition. Suppose that a is a limit action under the optimal policy π, and let $(a_i, y_i)_{i=1}^\tau$ be a history with positive probability. We show that if the agent plays a at every period after $(a_i, y_i)_{i=1}^\tau$ almost surely the belief μ_t reaches a region where no optimal policy prescribes a. By Lemma 5 this is enough to obtain the desired conclusion. Since a is not a uniform BN-E, then there is $p' \in \hat{\Theta}(a)$ such that if $\operatorname{supp} \nu \subseteq \mathcal{E}_a(p')$, then $a \notin A^m(\nu)$. We set $q = p'_a$ throughout this proof.

Claim 1. *There exists* $\varepsilon > 0$ *such that if*

$$\frac{\nu\left(\{p \in \Theta \colon \forall y \in \operatorname{supp} p_a^*, |p_a(y) - q(y)| < \varepsilon\}\right)}{1 - \nu\left(\{p \in \Theta \colon \forall y \in \operatorname{supp} p_a^*, |p_a(y) - q(y)| < \varepsilon\}\right)} > \frac{1-\varepsilon}{\varepsilon},$$

then $\pi(\nu) \neq a$.

Proof. Suppose that for every n there exists a ν_n such that

$$\frac{\nu_n\left(\{p \in \Theta \colon \forall y \in \operatorname{supp} p_a^*, |p_a(y) - q(y)| < 1/n\}\right)}{1 - \nu_n\left(\{p \in \Theta \colon \forall y \in \operatorname{supp} p_a^*, |p_a(y) - q(y)| < 1/n\}\right)} \geqslant \frac{1-1/n}{1/n}$$

and $a \in \pi(\nu_n)$. Because $\Delta(\Theta)$ is sequentially compact, $(\nu_n)_{n \in \mathbb{N}}$ has a converging subsequence $(\nu_{n_i})_{i \in \mathbb{N}} \to \nu^*$.

To show that this leads to a contradiction, define $G(\nu) = \max_{\tilde{\pi}} V(\tilde{\pi}, \nu) - \max_{\tilde{\pi}:\tilde{\pi}(\nu)=a} V(\tilde{\pi}, \nu)$. We claim that if $\operatorname{supp} \nu \subseteq \{p \in \Theta \colon \forall y \in \operatorname{supp} p_a^*, p_a(y) = q(y)\}$, then $G(\nu) > 0$. This is because the definition of q implies $\operatorname{supp} \nu \subseteq \mathcal{E}_a(p')$, so $a \notin A^m(\nu)$, and $\operatorname{supp} \nu \subseteq \mathcal{E}_a(p')$ implies that the experimentation value of a is 0.

Next note that as shown in Lemma 2, the space of policy functions endowed with the product topology is compact and $V(\cdot, \nu) - V(\cdot, \nu)$ is a continuous function of the policy. Since for every policy $\tilde{\pi}$, $\mathbb{E}_{p,\tilde{\pi}}\left[\sum_{t=1}^\infty \left[\beta^{t-1} u(a_t, y_t)\right]\right]$ is continuous in p, and $V(\tilde{\pi}, \cdot)$ is continuous in ν, so from the Maximum Theorem G is continuous. So if the claim is false, $\nu^* \left(\{p \in \Theta \colon \forall y \in \operatorname{supp} p_a^*, p_a(y) = q(y)\}\right) = 1$ and $G(\nu^*) = 0$, which would imply that $a \in \pi(\nu^*)$, a contradiction. ■

In what follows, fix such an ε. Also, fix an outcome $y^0 \in \operatorname{supp} p_a^*$, and let $\tilde{f}_t$ be the empirical frequency of the other $|\operatorname{supp} p_a^*| - 1$ outcomes in the support of p_a^*. Denote by $\tilde{p}_a^*$ the true probabilities of the same $|\operatorname{supp} p_a^*| - 1$ outcomes.

Claim 2. *$\tilde{f}_t \cdot t - \tilde{p}_a^* t$ is a $|\operatorname{supp} p_a^*| - 1$ dimensional random walk under the distribution $\tilde{p}_a^*$, and the covariance matrix of its increments is nonsingular.*

Proof. Let $y \in \operatorname{supp} p_a^* \backslash \{y^0\}$. The increment of the y dimension at time $t+1$ is equal to

$$\tilde{f}_{t+1}(y) \cdot (t+1) - p_a^*(y) \cdot (t+1) - \tilde{f}_t(y) \cdot t - p_a^*(y) \cdot t = \mathbf{1}_{y_{t+1}=y} - p_a^*(y)$$

and has expected value 0. Therefore, $\tilde{f}_t \cdot t - \tilde{p}_a^* t$ is a $|\operatorname{supp} p_a^*| - 1$ dimensional random walk.

The covariance matrix for the increments is given by $\Sigma_{y,y'} = -2\tilde{p}_a^*(y)\tilde{p}_a^*(y')$ if $y \neq y'$ and $2\tilde{p}_a^*(y)(1 - \tilde{p}_a^*(y))$ if $y = y'$.[28] If we let D be the identity matrix in part M35 of Theorem 2.3 of Berman and Plemmons (1994), for every $y' \in Y$, we have that

$$2\tilde{p}_a^*(y')(1 - \tilde{p}_a^*(y')) = 2\tilde{p}_a^*(y') \sum_{y \neq y'} \tilde{p}_a^*(y) > 2\tilde{p}_a^*(y') \sum_{y \neq y', y^0} \tilde{p}_a^*(y)$$

so the matrix is diagonal dominant and therefore not singular. ■

By the Central Limit Theorem $(\tilde{f}_t - \tilde{p}_a^*)\sqrt{t}$ converges to a Normal random variable with mean 0 and covariance matrix $\Sigma_{y,y'}$. Let $F_t = B_{\frac{||q - p_a^*||/K'}{\sqrt{t}}}\left(\tilde{p}_a^* + \frac{1}{\sqrt{t}}(q - p_a^*)\right)$. We have that

$$\mathbb{P}\left[\tilde{f}_t \in F_t\right] = \mathbb{P}\left[\sqrt{t}(\tilde{f}_t - \tilde{p}_a^*) \in B_{||q - p_a^*||/K'}(q - p_a^*)\right]$$

Taking the limit $t \to \infty$ yields that

$$\lim_{t \to \infty} \mathbb{P}\left[\tilde{f}_t \in F_t\right] = \mathbb{P}\left[\tilde{Z} \in B_{||q - p_a^*||/K'}(q - p_a^*)\right]$$

where $\tilde{Z}$ is a random variable that is normally distributed with mean $\vec{0}$ and covariance matrix

[28] This is verified in Claim 3 of the Online Appendix.

$\Sigma_{y,y'}$. Thus if we let E_t denote the event $f_t \in F_t$, it follows that $\sum_{t=1}^{\infty} \mathbb{P}[E_t] = \infty$. Moreover,

$$\begin{aligned}
\liminf_{t\to\infty} \frac{\sum_{s=1}^{t} \sum_{r=1}^{t} \mathbb{P}[E_s \text{ and } E_t]}{\left(\sum_{s=1}^{t} \mathbb{P}[E_s]\right)^2} &= \liminf_{t\to\infty} \frac{\frac{1}{t^2} \sum_{s=1}^{t} \sum_{r=1}^{t} \mathbb{P}[E_s \text{ and } E_r]}{\left(\frac{1}{t} \sum_{t=1}^{\infty} \mathbb{P}[E_t]\right)^2} \\
&\leqslant \liminf_{t\to\infty} \frac{\frac{1}{t^2} \sum_{s=1}^{t} \sum_{r=1}^{t} \mathbb{P}[E_r]}{\left(\frac{1}{t} \sum_{s=1}^{t} \mathbb{P}[E_s]\right)^2} = \liminf_{t\to\infty} \frac{\frac{1}{t} \sum_{r=1}^{t} \mathbb{P}[E_r]}{\left(\frac{1}{t} \sum_{s=1}^{t} \mathbb{P}[E_s]\right)^2} \\
&= \frac{1}{\lim_{t\to\infty} \mathbb{P}[E_t]} = \frac{1}{\mathbb{P}\left[\tilde{Z} \in B_{||q-p_a^*||/K'}\left(q - p_a^*\right)\right]} .
\end{aligned}$$

It then follows from the Kochen-Stone lemma (see Kochen and Stone (1964) or Exercise 2.3.20 in Durrett (2008)) that

$$\mathbb{P}\left[\bigcap_{t=1}^{\infty} \bigcup_{s=t}^{\infty} E_s\right] \geqslant \mathbb{P}\left[\tilde{Z} \in B_{||q-p_a^*||/K'}\left(q - p_a^*\right)\right] > 0 .$$

The event $\bigcap_{t=1}^{\infty} \bigcup_{s=t}^{\infty} E_s$ is invariant under finite permutations of the increments $\left(\mathbf{1}_{y_t=y^1}, \ldots, \mathbf{1}_{y_t=y^{|\operatorname{supp} p_a^*|-1}} - p_a^*\right)$ with different time indices, so the Hewitt–Savage zero–one law (see, e.g., Theorem 8.4.6 in Dudley, 2018) implies that the probability of the event $\bigcap_{t=1}^{\infty} \bigcup_{s=t}^{\infty} E_s$ is zero or one, and since it is strictly positive it must equal one.

This implies that $f_t \in F_t$ infinitely often with probability 1. So, by Proposition 1 the agent will eventually take an action different from a. ■

Proof of Theorem 2. *If.* Consider a uniformly strict BN-E a, an optimal policy π and $\kappa \in (0,1)$. By Lemma 3, there exists an ε such that if $\nu(\hat{\Theta}^{\varepsilon}(a)) \geqslant 1 - \varepsilon$, then $\pi(\nu) = a$.

Recall that for every $l \in (0,1)$, the function $f_l : P \times P \to \bar{\mathbb{R}}$ is defined by

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

By Lemma 6, since $\hat{\Theta}^{\varepsilon}(a)$ is compact by Lemma 1, and since f_l is lower semicontinuous in its first argument, there exists $\varepsilon' \in (0, \varepsilon)$ such that $\bar{p} \in \hat{\Theta}^{\varepsilon'}(a)$ implies that $f_l(\bar{p}, p') < 1$ for all p' with $p' \notin \hat{\Theta}^{\varepsilon}(a)$. Let $K = \left(\frac{\varepsilon}{1-\varepsilon}\right)^l$. Then

$$\begin{aligned}
&\left(\frac{1 - \nu\left(\hat{\Theta}^{\varepsilon}(a)\right)}{\nu\left(\hat{\Theta}^{\varepsilon'}(a)\right)}\right)^l < K \implies \frac{1 - \nu\left(\hat{\Theta}^{\varepsilon}(a)\right)}{\nu\left(\hat{\Theta}^{\varepsilon}(a)\right)} < \frac{\varepsilon}{1-\varepsilon} \\
&\implies \nu\left(\hat{\Theta}^{\varepsilon}(a)\right) > 1 - \varepsilon \implies \pi(\nu) = a.
\end{aligned}$$

Let $\bar{\varepsilon}$ be such that $\nu\left(\hat{\Theta}^{\bar{\varepsilon}}(a)\right) > 1-\bar{\varepsilon}$ implies that

$$\left(\frac{1-\nu\left(\hat{\Theta}^{\varepsilon}(a)\right)}{\nu\left(\hat{\Theta}^{\varepsilon}(a)\right)}\right)^{l} < \frac{K\left(1-\kappa\right)}{n}.$$

Then if the agent starts with a belief ν_0 with $\nu_0(\hat{\Theta}^{\varepsilon}(a)) > \bar{\varepsilon}$, $A\left(\nu_0\right) = \{a\}$. Moreover, by Lemma 4, Dubins' upcrossing inequality, the compactness of $\hat{\Theta}^{\varepsilon}(a)$ guaranteed by Lemma 1, and the union bound, there is a probability $(1-\kappa)$ that the positive supermartingale

$$\left(\frac{1-\nu'_t\left(\hat{\Theta}^{\varepsilon}(a)\right)}{\nu'_t\left(\hat{\Theta}^{\varepsilon}(a)\right)}\right)^{l}$$

never rises above K, so the action played is always a, and $\bar{\varepsilon}$ satisfies the requirement of the statement.

Only if. If a is not a uniformly strict BN-E, there exists $p \in \hat{\Theta}(a)$ and $b \neq a$ such that $\{b\} \in A^{m}\left(\delta_p\right)$. But then if we let $\nu = \delta_p$ we have that $\nu\left(\hat{\Theta}(a)\right) = 1$. Moreover, there exists a policy π that prescribes b at belief ν, so that the agent will never update their belief and will play b forever. ■

Proof of Theorem 3. $(i) \Rightarrow (ii)$ Immediately follows by Theorem 2.

$(ii) \Rightarrow (i)$ We prove the statement by contraposition. Suppose that a is not a uniformly strict BN-E, and let $\nu \in \Delta\left(\Theta\right)$, $\varepsilon > 0$. We construct an initial belief ν_ε that is ε close to ν but such that the actions do not converge to a.

Since a is not a uniformly strict BN-E, there exists $\hat{p} \in \hat{\Theta}(a)$ with $\{a\} \neq A^{m}\left(\delta_{\hat{p}}\right)$. Let $\left(C_{\varepsilon,i}\right)_{i=1}^{n}$ be a finite collection of open balls of radius ε in $\Delta\left(\Delta\left(Y\right)^{A}\right)$ that covers $\hat{\Theta}(a)$ and such that for each $C_{\varepsilon,i} \cap \hat{\Theta}(a) \neq \varnothing$. For every $C_{\varepsilon,i}$, choose $q_{\varepsilon,i} \in C_{\varepsilon,i} \backslash \hat{\Theta}(a)$ whose existence follows from the assumption of the theorem.

Define $\Phi_\varepsilon : \Theta \rightarrow 2^{\Theta}$ as

$$\Phi_\varepsilon\left(p\right) = \begin{cases} \{q_{\varepsilon,i} : p \in C_{\varepsilon,i}\} & \text{if } p \in C_{\varepsilon,i} \text{ for some } i \\ \{p\} & \text{otherwise.} \end{cases}$$

The correspondence Φ_ε is Borel measurable, nonempty, and closed valued, so it has a measurable selection ϕ_ε by the Kuratowski Selection Theorem (see, e.g., Theorem 18.13 in Aliprantis and Border, 2013). Define $\bar{\nu}_\varepsilon\left(C\right) = \nu\left(\phi_\varepsilon^{-1}\left(C\right)\right)$. Because the problem is rich, there is $p' \in \Theta \cap B_\varepsilon(\hat{p})$ such that $H(p'_a, p^*_a) < \min_{p \in \operatorname{supp} \bar{\nu}_\varepsilon} H(p_a, p^*_a)$ and $a \notin A^m(\delta_{p'})$. Set $\nu_\varepsilon = \varepsilon \delta_{p'} + (1-\varepsilon)\,\bar{\nu}_\varepsilon$. Then $\nu_\varepsilon \to \nu$, but $\underset{p' \in \operatorname{supp} \nu_\varepsilon}{\operatorname{argmin}} H\left(p^*_a, p'_a\right) = \{\hat{p}\}$, so by Theorem 1, the probability of converging to a starting from belief ν_ε is 0. ■

Proof of Theorem 4. Since the agent believes that actions do not change the outcome distribution, every $p \in \Theta$ can be identified with an element of $\Delta(Y)$, and every belief $\nu \in \Delta(\Theta)$ can be identified with an element of $\Delta(\Delta(Y))$.

Consider a uniformly strict BN-E a. By Lemma 1, $\Delta(\hat{\Theta}(a))$ is compact. For every $\bar{\varepsilon} > 0$ and $q \in \Delta(Y)$ let $Q_{\bar{\varepsilon}}(q) = Q_{\bar{\varepsilon}, \mu_{0,a}}(q)$. By Theorem 2, there exists $\varepsilon' > 0$ such that if $\varepsilon' > \varepsilon$ and $\nu\left(\operatorname{cl}\left(Q_\varepsilon\left(p^*_a\right)\right)\right) > 1 - \varepsilon$ implies $A^m\left(\nu\right) = \{a\}$ the probability of playing a forever starting from belief ν is larger than $1/2$. By the Maximum Theorem, the correspondence Q_ε is upper-hemicontinuous, so there is a sequence of outcomes y^t with corresponding empirical frequency $\hat{p}_t(y) = \frac{1}{t}\sum_{i=1}^t \mathbf{1}_{y_i = y}$ sufficiently close to p^*_a to have

$$\hat{q} \in Q_{\varepsilon'/2}\left(\hat{p}_t\right), q \in Q_{\varepsilon'/2}\left(p^*_a\right) \implies ||\hat{q} - q|| < \varepsilon/2.$$

This implies $Q_{\varepsilon'/2}\left(\hat{p}_t\right) \subseteq Q_{\varepsilon'}\left(p^*_a\right)$ from the triangle inequality. Thus by Lemma 8 there is a time T such that for all $t' > T$, if the empirical frequency is $\hat{p}_{t'} = \hat{p}_t$, the agent assigns a relative probability higher than K to an ε' ball around p^*_a:

$$\frac{\mu_{t'}(Q_{\varepsilon'}\left(p^*_a\right))}{1 - \mu_{t'}(Q_{\varepsilon'}\left(p^*_a\right))} \geqslant \frac{\mu_{t'}(Q_{\varepsilon'/2}\left(\hat{p}_t\right))}{1 - \mu_{t'}(Q_{\varepsilon'}\left(p^*_a\right))} > \frac{K}{2}.$$

Replicating y^t sufficiently many times yields a sequence $y^{t'}$ with empirical frequency $\hat{p}_{t'} = \hat{p}_t$ and $t' > T$. Since p^*_a is absolutely continuous with respect to $p^*_{a'}$ for all $a' \in A$, this sequence of outcomes has positive probability, and after it occurs the agent plays a. By Lemma 4 and the law of iterated expectations, conditional on a being played $\left(\frac{1-\mu_{t'}(Q_\varepsilon(p^*_a))}{\mu_{t'}(Q_{\varepsilon'/2}(\hat{p}_t))}\right)^l$ is a positive supermartingale. Then by Dubins' upcrossing inequality, there is positive probability that this positive supermartingale never rises above $1/K^l$, so a is played forever. ■

Proof of Theorem 5. Let b be a weakly identified strict BN-E. Then there is $\nu \in \Delta(\hat{\Theta}(b))$ with $b = A^m(\nu)$. Since b is a strict BN-E, and the agent believes the outcome distributions are independent across actions, we can let $\nu = \delta_p$ where $p_b = \operatorname{argmax}_{p'_b : p' \in \Theta} \mathbb{E}_{p'_b}\left[u(b, y)\right]$, and $p_a = \operatorname{argmin}_{p'_a : p' \in \Theta} \mathbb{E}_{p'_a}\left[u(a, y)\right]$ for $a \in A \backslash \{b\}$. Let $\{y\left(b\right)_i\}_{i=1}^\infty$ be a sequence of outcomes such that the empirical frequency $\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{y_i = y}$ is converging to p_b. By Lemma 8, for every $\varepsilon > 0$, there exists K_ε such that for all $t > K_\varepsilon$, $\mu_{0,b}\left(B_\varepsilon(p_b) \mid y\left(b\right)^t\right) > 1 - \varepsilon$.

Because $\{b\} = A^m(\nu)$, there are $\bar{\beta} \in (0,1)$ and $(\varepsilon_a)_{a \in A} \in \mathbb{R}_+^A$ such that if (i) $\beta > \bar{\beta}$ and (ii) the belief $\bar{\nu}$ is such that $\bar{\nu}_b \in \{\mu_{0,b}(\cdot \mid y(b)_t) : 0 \leqslant t \leqslant K_{\varepsilon_b}\} \bigcup \{\nu'_b : \nu'_b(B_\varepsilon(p_b)) > 1 - \varepsilon\}$, and for all $a' \neq b$, $\nu_{a'}\left(B_{\varepsilon_{a'}}(p_{a'})\right) > 1 - \varepsilon_{a'}$, then b has the highest Gittins index. For each $\beta > \bar{\beta}$, let $\varepsilon_\beta < \varepsilon$ be such that if $\bar{\nu}_b\left(B_{\varepsilon_\beta}(p(b))\right) > (1 - \varepsilon_\beta)$ then the probability of converging to play action a is larger than $\frac{1}{2}$ under any optimal policy given the discount factor β, whose existence is guaranteed by Lemma 14 and the fact that b is weakly identified.

For every $a \neq b$, let $\bar{n}_a \geqslant n_a$ and $\{y(a)_i\}_{i=1}^{n_a}$ be a sequence of outcomes such that the empirical frequency $\hat{p}_{n_a}(a)$ converges to p_a. By Lemma 8, for every $a \neq b$ there is a finite n_a such that after n_a observations $\nu_a(B_{\varepsilon_a}(p_a) \mid \hat{p}_{n_a}) > 1 - \varepsilon_a$. Finally, let $n_b = K_{\varepsilon_\beta}$. Then the array $\left(\{y(a)_i\}_{i=1}^{n_a}\right)_{a \in A}$ has positive probability, so 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 Corollary 2. If a is quasi-dominated, there exists ε such that if $\nu_a(\{q : ||q - p_a|| \leqslant \varepsilon\}) > 1 - \varepsilon$ implies $\pi(\nu) \neq a$. Suppose by way of contradiction that a is played infinitely many times. Then by the last part of the proof of Theorem 1, since the the problem is a subjective bandit, if action a is played infinitely many times then there is t such that $\mu_a(\{q : ||q - p_a|| \leqslant \varepsilon\} | (a^t, y^t)) > 1 - \varepsilon$, so the agent switches to another action b. Since while playing an action different from a the agent does not update μ_a, $\mu_a(\{q : ||q - p_a|| \leqslant \varepsilon\} | (a^\tau, y^\tau)) > 1 - \varepsilon$ for all $\tau > t$, so they will not switch to a anymore, a contradiction. ■

Proof of Theorem 6. We prove the statement for $\bar{a}$, the proof for $\underline{a}$ is analogous. Denote the optimal policy used by the agent as π. Since the environment is strongly supermodular, every class of observationally equivalent outcome distributions under action $\bar{a}$ is a singleton, so $\bar{a}$ is a uniformly strict BN-E. Theorem 2 and the strong supermodularity of the environment then imply there is $\bar{p} \in \Theta$ and $K \in (0,1)$ such that if $\nu(\{p : p > \bar{p}\}) > K$, then the probability that a is 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 $t \leqslant n(b)$, then $\mu_t(\{p : p > \bar{p}\} \mid (a^t, y^t)) \geqslant K$.

Consider the event E that for all $b \in A$ and $t \leqslant n(b)$, $x_{t,b} = \bar{y}$. This event has strictly positive probability $\mathbb{P}_\pi[E]$. Moreover, if E realizes, after some $\hat{T} \leqslant \sum_{b \neq \bar{a}}(n(b) - 1) + 1$, the policy of the agent prescribes action $\bar{a}$. Therefore, after $\hat{T} + n(\bar{a})$, for all $\tau \leqslant \hat{T} + n(\bar{a})$, and for all $y \in Y, \mathbb{P}\left[x_{\tau,\bar{a}} = y|E\right] = \mathbb{P}\left[x_{\tau,\bar{a}} = y\right]$. Therefore, by Theorem 2 the probability of converging to $\bar{a}$ is at least $\frac{\mathbb{P}_\pi[E]}{2}$. ■

A.5 Action Frequencies and Mixed Equilibria

By Theorem 1, if action a is not a uniform BN-E, the agent will use a different action b infinitely often. We can the use a result from of Esponda, Pouzo, and Yamamoto (2019) to show that if action b's outcome distribution does not induce a as a myopic best reply, the agent will spend a nontrivial fraction of time using actions different from a. For every $a \in A$, let $\underline{\Theta}(a) = \left\{p \in \hat{\Theta}(a) : a \notin A^m(\delta_p)\right\}$. Let $C_a \subseteq \Delta(\Theta)$ be the largest convex set such that (1) it contains all ν with $\operatorname{supp}\nu = \Theta(a) \backslash \underline{\Theta}(a)$, and (2) $a \in A^m(\nu)$ for all $\nu \in C_a$. That is, C_a contains all the beliefs supported on the "good" KL minimizers for action a that induce a as a best reply, as well as the beliefs around them that still support a as a myopic best reply. Also, let $A_a = \{b : \exists \nu \in C_a, b \in A^m(\nu)\}$.

Corollary 3. *Let $\beta = 0$, and suppose $a \in A$ is a non-uniform BN-E. If there is $\bar{p} \in \underline{\Theta}(a)$ such that $(p_b^*, \bar{p}_b) < H(p_b^*, \hat{p}_b)$ for all $b \in A_a$ and $\hat{p} \in \Theta \backslash \{\bar{p}\}$, then* $\liminf \frac{1_{a_t=a}}{t} \neq 1$ *a.s.*

Proof of Corollary 3. Let $\varepsilon > 0$ be such that if $||\bar{p} - p|| \leqslant \varepsilon$, then $a \notin A^m(\delta_p)$. Let $\underline{\Theta}^\varepsilon(a) = \{p \in \Theta : ||\bar{p} - p|| \leqslant \varepsilon\}$. By assumption, there exists $\varepsilon' > 0$ such that $\hat{\Theta}(\alpha) \subseteq \underline{\Theta}^\varepsilon(a)$ for all $\alpha \in \Delta(A)$ such that $||\alpha - a|| < \varepsilon'$, $\operatorname{supp}\alpha = A_a \bigcup\{a\}$. Suppose by way of contradiction that $\liminf \frac{1_{a_t=a}}{t} = 1$. Let $W_\tau(\alpha) \subseteq \Delta(A)^{[\tau,\infty)}$ be the set of all differentiable functions $\gamma : [\tau, \infty) \to \Delta(A)$ such that

$$\frac{\partial \gamma_t}{\partial t} \in \Delta\left(A^m\left(\Delta\left(\hat{\Theta}(\gamma_t)\right)\right)\right) - \gamma_t$$

and $\gamma_0 = \alpha$. Define the random variable $\hat{\alpha}_t$ to be the empirical frequency of actions up to time t, i.e., $\hat{\alpha}_t(b) = \frac{1_{a_t=b}}{t}$ for all $b \in A$. For every $\tau \in [t, t+1]$ let $\hat{\alpha}_\tau(b) = \hat{\alpha}_t(b)(\tau - t) + \hat{\alpha}_{t+1}(b)(t+1-t)$. From the convergence result (Theorem 2) of Esponda, Pouzo, and Yamamoto (2019), for all $T > 0$ $\lim_{t\to\infty}\inf_{\gamma_t \in W_t(\hat{\alpha}_t)}\sup_{0 \leqslant s \leqslant T}||\hat{\alpha}_{t+s}(a) - \gamma_{t+s}(a)|| = 0$ a.s. By Theorem 1, for all $t' \in \mathbb{N}$ almost surely there is a $\hat{t} \geqslant t'$ such that $\mu_{\hat{t}} \notin C_a$. But then, since the frequency of action a decreases in a ball of size ε outside C_a, for all $\gamma \in W_{\hat{t}}(\hat{\alpha}_{\hat{t}})$, we have $||\gamma_{\hat{t}+\varepsilon'\hat{t}}(a) - 1|| > \varepsilon'$ and $\lim_{s\to\infty}\hat{\alpha}_{\hat{t}+s} = a$, a contradiction. ■

There are two reasons that multiple actions can be played with positive probability in a BN-E: Either every action played can be justified with the same belief over the KL minimizers, or different beliefs are needed to justify some of them. The first case requires the agent to be indifferent between the different actions, so here the BN-E cannot be uniformly strict. However, signals that take the form of payoff perturbations can allow us to obtain such equilibria as the limit of uniformly strict Berk-Nash equilibria, and the associated purification can be uniformly stable and positively attractive.

Formally, for every $\alpha \in \Delta(A)$ and $p \in \Theta$, let

$$H_\alpha(p^*, p) = \sum_{b \in A} \alpha(b) p_b^*(y) \log p_b(y) \text{ and } \hat{\Theta}(\alpha) = \operatorname*{argmin}_{p \in \Theta} H_\alpha(p^*, p).$$

Definition 14 (Strongly Uniform Mixed BN-E)**.** The mixed action $\alpha \in \Delta(A)$ is a *strongly uniform mixed BN-E* if all actions $a \in \operatorname{supp} \alpha$ are myopically optimal for all $\theta \in \hat{\Theta}(\alpha)$.

Given a problem (A, Y, p^*, u, Θ) without signals, a problem with signals $(A, Y, S, \zeta, \tilde{p}^*, \tilde{u}, \tilde{\Theta})$ is its (ε, v) *perturbation*, $\varepsilon \in \mathbb{R}_+, v : A \times Y \times S \to \mathbb{R}$, if (i) $\tilde{u}(a, y, s) = u(a, y) + \varepsilon v(a, y, s)$, (ii) $\tilde{p}_{a,s}^*(y) = p_a^*(y)$ and (iii) $\tilde{\Theta} = \{\tilde{p} : \exists p \in \Theta, \tilde{p}_{a,s}(y) = p_a(y), \forall (a, y, s) \in A \times Y \times S\}$.

Corollary 4. *If α is a strongly uniform mixed BN-E in (A, Y, p^*, u, Θ), there is a sequence of strategies $(\sigma_n)_{n \in \mathbb{N}}$ such that each $\sigma_{1/n}$ is a uniformly stable BN-E of a $(1/n)$-perturbation of (A, Y, p^*, u, Θ) and $\lim_{n \to \infty} \zeta(s : \sigma_n(s) = a) = \alpha(a)$ for all $a \in A$. If (A, Y, p^*, u, Θ) is subjectively exogenous and p^* has full support, there are positively attractive $\sigma_{1/n}$.*

The proof is in Section B.2 of the Online Appendix.

References

Aliprantis, C. and K. Border (2013). *Infinite Dimensional Analysis: A Hitchhiker's Guide.* Berlin. Springer-Verlag.

Arrow, K. and J. Green (1973). "Notes on Expectations Equilibria in Bayesian Settings". Working Paper No. 33, Stanford University.

Bell, A. et al. (2019). "Do tax cuts produce more Einsteins? The impacts of financial incentives versus exposure to innovation on the supply of inventors". *Journal of the European Economic Association.*

Benaïm, M. and M. W. Hirsch (1999). "Mixed Equilibria and Dynamical Systems Arising from Fictitious Play in Perturbed Games". *Games and Economic Behavior* 29, pp. 36–72.

Berk, R. H. (1966). "Limiting Behavior of Posterior Distributions when the Model is Incorrect". *The Annals of Mathematical Statistics* 37, pp. 51–58.

Berman, A. and R. J. Plemmons (1994). *Nonnegative Matrices in the Mathematical Sciences*. SIAM.

Bohren, J. A. (2016). "Informational Herding with Model Misspecification". *Journal of Economic Theory* 163, pp. 222–247.

Bohren, J. A. and D. Hauser (2020). "Learning with Model Misspecification: Characterization and Robustness".

Bray, M. (1982). "Learning, estimation, and the stability of rational expectations". *Journal of economic theory* 26, pp. 318–339.

Bray, M. M. and N. E. Savin (1986). "Rational expectations equilibria, learning, and model specification". *Econometrica*, pp. 1129–1160.

Cho, I.-K. and K. Kasa (2015). "Learning and model validation". *The Review of Economic Studies* 82, pp. 45–82.

— (2017). "Gresham's Law of Model Averaging". *American Economic Review* 107, pp. 3589–3616.

Diaconis, P. and D. Freedman (1990). "On the Uniform Consistency of Bayes Estimates for Multinomial Probabilities". *The Annals of Statistics* 18, pp. 1317–1327.

Dudley, R. M. (2018). *Real Analysis and Probability*. Chapman and Hall/CRC.

Durrett, R. (2008). *Probability Models for DNA Sequence Evolution*. Springer.

Eliaz, K. and R. Spiegler (2018). "A Model of Competing Narratives". arXiv: 1811.04232.

Esponda, I. and D. Pouzo (2016). "Berk–Nash equilibrium: A Framework for Modeling Agents with Misspecified Models". *Econometrica* 84, pp. 1093–1130.

— (2019). "Equilibrium in Misspecified Markov Decision Processes". arXiv: 1502.06901.

Esponda, I., D. Pouzo, and Y. Yamamoto (2019). "Asymptotic Behavior of Bayesian Learners with Misspecified Models". arXiv: 1904.08551.

Frick, M., R. Iijima, and Y. Ishii (2019). "Misinterpreting Others and the Fragility of Social Learning".

— (2020). "Stability and Robustness in Misspecified Learning Models".

Fudenberg, D. and K. He (2017). "Player-compatible Equilibrium". arXiv: 1712.08954.

— (2018). "Learning and type compatibility in signaling games". *Econometrica* 86, pp. 1215–1255.

— (2020). "Payoff information and learning in signaling games". *Games and Economic Behavior* 120, pp. 96–120.

Fudenberg, D., K. He, and L. A. Imhof (2017). "Bayesian posteriors for arbitrarily rare events". *Proceedings of the National Academy of Sciences* 114, pp. 4925–4929.

Fudenberg, D. and D. M. Kreps (1993). “Learning Mixed Equilibria”. *Games and Economic Behavior* 5, pp. 320–367.

Fudenberg, D., G. Lanzani, and P. Strack (2020). “Uniform consistency of beliefs without full support”.

Fudenberg, D. and D. K. Levine (1992). “Maintaining a Reputation When Strategies are Imperfectly Observed”. *Review of Economic Studies* 59, pp. 561–581.

Fudenberg, D., G. Romanyuk, and P. Strack (2017). “Active Learning with a Misspecified Prior”. *Theoretical Economics* 12, pp. 1155–1189.

Gagnon-Bartsch, T. M. (2016). “Taste Projection in Models of Social Learning”. PhD thesis.

Gibbons, R., M. LiCalzi, and M. Warglien (2019). *What situation is this? Coarse cognition and behavior over a space of games*. Tech. rep. Department of Management, Università Ca’Foscari Venezia.

He, K. (2019). “Mislearning from Censored Data: The Gambler’s Fallacy in Optimal-Stopping Problems”. arXiv: 1803.08170.

Heidhues, P., B. Kőszegi, and P. Strack (2018a). “Convergence in Misspecified Learning Models with Endogenous Actions”. *Available at SSRN 3312968*.

— (2018b). “Unrealistic Expectations and Misguided Learning”. *Econometrica* 86, pp. 1159–1214.

— (2019). “Overconfidence and Prejudice”. arXiv: 1909.08497.

Jehiel, P. (2018). “Investment strategy and selection bias: An equilibrium perspective on overoptimism”. *American Economic Review* 108, pp. 1582–97.

Kagel, J. H. and D. Levin (1986). “The winner’s curse and public information in common value auctions”. *The American economic review*, pp. 894–920.

Kochen, S. and C. Stone (1964). “A Note on the Borel-Cantelli Lemma”. *Illinois Journal of Mathematics* 8, pp. 248–251.

Levy, G., R. Razin, and A. Young (2020). “Misspecified Politics and the Recurrence of Populism”. Working Paper.

Mailath, G. J. and L. Samuelson (2019). “Learning under Diverse World Views: Model-Based Inference”.

Molavi, P. (2019). “Macroeconomics with Learning and Misspecification: A General Theory and Applications”.

Morrison, W. and D. Taubinsky (2019). “Rules of Thumb and Attention Elasticities: Evidence from Under-and Overreaction to Taxes”. NBER Working Paper.

Neveu, J. (1975). *Discrete-parameter martingales*. Amsterdam. North-Holland.

Nyarko, Y. (1991). “Learning in Mis-specified Models and the Possibility of Cycles”. *Journal of Economic Theory* 55, pp. 416–427.

Parthasarathy, K. R. (2005). *Probability Measures on Metric Spaces*. American Mathematical Soc.

Rabin, M. and J. L. Schrag (1999). “First impressions matter: A model of confirmatory bias”. *The Quarterly Journal of Economics* 114, pp. 37–82.

Rabin, M. and D. Vayanos (2010). “The gambler’s and hot-hand fallacies: Theory and applications”. *The Review of Economic Studies* 77, pp. 730–778.

Rees-Jones, A. and D. Taubinsky (2016). “Measuring “Schmeduling””. NBER Working Paper.

Tversky, A. and D. Kahneman (1973). “Availability: A heuristic for judging frequency and probability”. *Cognitive Psychology* 5, pp. 207–232.

B Online Appendix

Lemma 1. *For every $a \in A$ and $\varepsilon > 0$, $\hat{\Theta}(a)$ defined in (1), $\hat{\Theta}_a(a)$, $\hat{\Theta}^{\varepsilon}(a)$ defined in (2), and $\Delta(\hat{\Theta}(a))$ are compact.*

Proof. Compactness of $\hat{\Theta}(a)$ follows from Assumption 1 and Theorem 2.43 of Aliprantis and Border (2013). Since the projection map is continuous, and $\hat{\Theta}_a(a)$ is the projection of $\hat{\Theta}(a)$, $\hat{\Theta}_a(a)$ is compact as well. Since $\hat{\Theta}_a(a)$ is closed, it immediately follows that $\hat{\Theta}^{\varepsilon}(a)$ is closed as well, henceforth compact. Given the compactness and separability of $\hat{\Theta}(a)$, $\Delta(\hat{\Theta}(a))$ is compact by, e.g., Theorem 6.4 in Parthasarathy (2005). ■

Lemma 7. *Fix $q \in \Delta(Y)$ with $\operatorname{supp} q \subseteq \operatorname{supp} p_a^*$ and a compact set $C \subseteq \Delta(Y)$ that contains a $\tilde{q}$ with $\operatorname{supp} p_a^* \subseteq \operatorname{supp} \tilde{q}$. Then there exists a $K > 0$ such that for every $f' \in U_\varepsilon(q, p_a^*, \eta)$ with $\operatorname{supp} f' \subseteq \operatorname{supp} p_a^*$*

$$|\min_{q' \in C} H\left((1-\eta)p_a^* + \eta q, q'\right) - H\left((1-\eta)p_a^* + \eta q, q\right) - \min_{q' \in C} H\left(f', q'\right) + H\left(f', q\right)| \leqslant K\varepsilon\,.$$

Proof. Fix an arbitrary compact C that contains a $\tilde{q}$ with $\operatorname{supp} p_a^* \subseteq \operatorname{supp} \tilde{q}$. First, notice that by the Maximum Theorem,

$$\hat{C}(\eta, \varepsilon) := \bigcup_{f \in U_\varepsilon(q, p_a^*, \eta)} \operatorname*{argmin}_{q' \in C} H\left(f, q'\right)$$

is a compact-valued and upper-hemicontinuous correspondence. So, if we let

$$\hat{C} := \bigcup_{\varepsilon \in [0,1]} \bigcup_{\eta \in [0,1]} \hat{C}(\eta, \varepsilon),$$

there is a $K_1 > 0$ such that $\max_{y \in \operatorname{supp} p_a^*} \max_{q' \in \hat{C}} |\log q'(y)| < K_1$.

Then we have that for every $\eta \in [0,1]$, $\varepsilon > 0$, and $f \in U_\varepsilon(q, p_a^*, \eta)$:

$$\begin{aligned}
& |\min_{q' \in C} H\left((1-\eta)p_a^* + \eta q, q'\right) - H\left((1-\eta)p_a^* + \eta q, q\right) - \min_{q' \in C} H\left(f', q'\right) + H\left(f', q\right)| \\
\leqslant\ & |\min_{q' \in C} H\left((1-\eta)p_a^* + \eta q, q'\right) - \min_{q' \in C} H\left(f', q'\right)| + 2\varepsilon \max_{y \in \operatorname{supp} p_a^*} |\log q(y)| \\
\leqslant\ & |2K_1\varepsilon| + 2\varepsilon \max_{y \in \operatorname{supp} p_a^*} |\log q(y)|,
\end{aligned}$$

where the inequalities follows from $||f - (1-\eta)p_a^* + \eta q|| \leqslant \varepsilon$, and the definition of K_1. Thus $K :\ = 2(K_1 + \max_{y \in \operatorname{supp} p_a^*} |\log q(y)|) > 0$ satisfies the statement of the lemma. ■

Claim 3. *Let $\tilde{p}_a^* t$ and $\tilde{f}_t$ be defined as in the proof of Theorem 1. Then the covariance matrix for the increments of $\tilde{f}_t \cdot t - \tilde{p}_a^* t$ is given by $\Sigma_{y,y'} = -2\tilde{p}_a^*(y)\tilde{p}_a^*(y')$ if $y \neq y'$ and $2\tilde{p}_a^*(y)(1 - \tilde{p}_a^*(y))$ if $y = y'$.*

Proof. To see this, the covariance between 1_y and $1_{y'}$ is given by:

$$\begin{aligned}
& \tilde{p}_a^*(y)(1 - E(1_y))(0 - E(1_{y'})) + \tilde{p}_a^*(y')(0 - E(1_y))(1 - E(1_{y'})) \\
+ \quad & (1 - \tilde{p}_a^*(y') - \tilde{p}_a^*(y))(0 - E(1_y))(0 - E_{\tilde{p}_a^*}(1_{y'})) \\
= \quad & \tilde{p}_a^*(y)(1 - \tilde{p}_a^*(y))(-\tilde{p}_a^*(y')) + \tilde{p}_a^*(y')(-\tilde{p}_a^*(y))(1 - \tilde{p}_a^*(y')) \\
+ \quad & (1 - \tilde{p}_a^*(y') - \tilde{p}_a^*(y))(-\tilde{p}_a^*(y'))(-\tilde{p}_a^*(y)) \\
= \quad & -\tilde{p}_a^*(y)\tilde{p}_a^*(y')[(1 - \tilde{p}_a^*(y)) + (1 - \tilde{p}_a^*(y'))] + \tilde{p}_a^*(y')\tilde{p}_a^*(y)(1 - \tilde{p}_a^*(y') - \tilde{p}_a^*(y)) \\
= \quad & -\tilde{p}_a^*(y)\tilde{p}_a^*(y')[2 - \tilde{p}_a^*(y) - \tilde{p}_a^*(y')) + 1 + \tilde{p}_a^*(y') + \tilde{p}_a^*(y)] = -2\tilde{p}_a^*(y)\tilde{p}_a^*(y').
\end{aligned}$$

■

Computations for Example 1

The monopolist's payoff function if the valuations are uniformly distributed on $[0, 8]$ is $\mathbb{E}[u(a, y)|y \sim U([0, 8])] = \frac{8-a}{8}a$, so the unique optimal price from the set $\{3, 4, 5, 6, 7\}$ equals $a = 4$. If valuations are uniformly distributed on $[2, 10]$, the payoff function is $\mathbb{E}[u(a, y)|y \sim U([2, 10])] = \frac{10-a}{8}a$, so the unique optimal price is $a = 5$.

Let $p^L = (\frac{8-a}{8})_{a\in\{3,4,5,6,7\}}$ be the vector of conditional probabilities when the demand is low and $p^H = (\frac{10-a}{8})_{a\in\{3,4,5,6,7\}}$ be the vector of conditional probabilities when the demand is high. It is easy to check that the KL minimizers are given by

$$\hat{\Theta}(3) = \{p^H\} \quad \hat{\Theta}(4) = \{p^H\} \quad \hat{\Theta}(5) = \{p^L, p^H\} \quad \hat{\Theta}(6) = \{p^L\} \quad \hat{\Theta}(7) = \{p^L\}.$$

Thus, $a = 5$ is the only pure BN-E. Note that $a = 5$ is not a uniform BN-E, because at the low belief the optimal action is 4.

Example 6

Example 6. *This example shows that Theorem 1 does not hold without Assumption 1(ii). Let the action space be $\{a, b\}$, the outcome space be $Y = \{0, 1\}$, and suppose the agent correctly believes that the action has no impact on the outcome distribution, and that $p^* = \frac{1}{2}$.*

Assume that the agent assigns positive probabilities to the following countable set:

$$\left\{\frac{3}{4}\right\} \cup \left\{\frac{1}{4} - \frac{1}{n^2} : n \geqslant 3\right\},$$

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

Finally, 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 $\nu(\{3/4\}) > 1/2$ and it is strictly preferred to a if $\nu(\{3/4\}) < 1/3$. Then a is a BN-E but not a uniform BN-E, yet play can converge to it with positive probability from a prior μ_0 we specify below.

In the claim below we show that for every $n \in \mathbb{N}$ there exists a $l_n > 0$ such that

$$1 \leqslant p^*(1)\left(\frac{\frac{3}{4}}{p(n)(1)}\right)^{l_n} + p^*(0)\left(\frac{\frac{1}{4}}{p(n)(0)}\right)^{l_n}.$$

Then by Dubins' upcrossing inequality[29]*, for all K_1, and K_2 there exists $C_n \leqslant \frac{\frac{1}{n^2}}{2\sum_{n=3}^{\infty}\frac{1}{n^2}}$ such that if $\mu_0(p(n)) \leqslant C_n$ and $\mu_0\left(\frac{3}{4}\right) > \frac{1}{2}$, 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 $\mu_0\left(\frac{3}{4}\right) = 1 - \sum_{n=3}^{\infty} C_n > \frac{1}{2}$, $K_2 < \frac{1}{\sum_{n=3}^{\infty}\frac{1}{n^2}}$ 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

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

Claim 4. *Notice that the outcome distribution most favorable to action b and least favorable*

[29] See, e.g., page 27 of Neveu, 1975

to action a is $p(3) = 1/4 - 1/9 = 5/36$. Therefore, if $\nu_t(\{3/4\}) > 1/2$,

$$\begin{aligned}
&\int_{\Delta(Y)} \mathbb{E}_p[u(a,y)]\,d\nu(p) \geqslant \sum_{n=3}^{\infty} p(n)u(a,1)\nu(\{p(n)\}) + \frac{3}{4}u(a,1)\nu(\{3/4\}) \\
&\geqslant \frac{5}{36}u(a,1)(1-\nu(\{3/4\})) + \frac{3}{4}u(a,1)\nu(\{3/4\}) > 4/9
\end{aligned}$$

and

$$\begin{aligned}
\int_{\Delta(Y)} \mathbb{E}_p[u(b,y)]\,d\nu(p) &\leqslant \sum_{n=3}^{\infty}(1-p(n))u(b,0)\nu(\{p(n)\}) + \frac{1}{4}u(b,0)\nu(\{3/4\}) \\
&\leqslant \frac{31}{36}u(b,0)(1-\nu(\{3/4\})) + \frac{1}{4}u(b,0)\nu(\{3/4\}) < 4/9.
\end{aligned}$$

If $\nu_t(\{3/4\}) < 1/3$,

$$\begin{aligned}
\int_{\Delta(Y)} \mathbb{E}_p[u(a,y)]\,d\nu(p) &\leqslant \sum_{n=3}^{\infty} p(n)u(a,1)\nu(\{p(n)\}) + \frac{3}{4}u(a,1)\nu(\{3/4\}) \\
&\leqslant \frac{1}{4}u(a,1)(1-\nu(\{3/4\})) + \frac{3}{4}u(a,1)\nu(\{3/4\}) < \frac{5}{12}
\end{aligned}$$

and

$$\begin{aligned}
\int_{\Delta(Y)} \mathbb{E}_p[u(b,y)]\,d\nu(p) &\geqslant \sum_{n=3}^{\infty}(1-p(n))u(b,0)\nu(\{p(n)\}) + \frac{1}{4}u(b,0)\nu(\{3/4\}) \\
&\geqslant \frac{3}{4}u(b,0)(1-\nu(\{3/4\})) + \frac{1}{4}u(b,0)\nu(\{3/4\}) = \frac{7}{15}.
\end{aligned}$$

Finally, notice that

$$\begin{aligned}
1 &\leqslant p^*(1)\left(\frac{\frac{3}{4}}{p(n)(1)}\right)^{l_n} + p^*(0)\left(\frac{\frac{1}{4}}{p(n)(0)}\right)^{l_n} \\
&= \frac{1}{2}\left(\frac{\frac{3}{4}}{\frac{1}{4}-\frac{1}{n^2}}\right)^{l_n} + \frac{1}{2}\left(\frac{\frac{1}{4}}{\frac{3}{4}+\frac{1}{n^2}}\right)^{l_n}
\end{aligned}$$

where

$$l_n = \frac{\log\left(1 - \frac{1}{\frac{4}{n^2}+3}\right)}{\log\left(\frac{1}{1-\frac{4}{n^2}}\right) + \log 3} > 0.$$

B.1 The role of Assumption 1(i)

All results in the paper except the non-myopic part of Theorem 1 continue to hold under a weaker version of Assumption 1(i):

Assumption 1(i′) 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 $p'_a(y) > 0$.

Assumption 1(i′) implies that the support of the belief does not change after a finite number of observations. This is the only consequence of Assumption 1(i) that is used in any of the proofs, except for establishing Claim 1 in the proof of Theorem 1 when the agent is not myopic.[30]

The next example shows that without Assumption 1(i′), limit points need not be BN-E.

Example 7 (Role of Assumption 1(i′))**.** *Suppose there are two actions a and b, and two outcomes $Y = \{0, 1\}$, and let $u(a, 0) = u(b, 1) = 1 - u(a, 1) = 1 - u(b, 0)$. Identify the elements of $\Delta(Y)$ with the probability they assign to outcome 1, and 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, and that $\Theta = \{\frac{1}{3}, 1\}$. Here b is the unique BN-E, and it is uniformly strict. However, if the prior assigns sufficiently high probability to $1/3$, 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 1 to $p = 1/3$ and the action converges to a.*

When we weaken Assumption 1(i) to (i′) and allow the supports the various outcome distributions to differ, we need to generalize the definition of observational equivalence as follows:

Definition 15. Two outcome distributions p and p' are *observationally equivalent under action a* if $p_a(y) = p'_a(y)$ for all $y \in \operatorname{supp} p^*_a$.

Thus we now say that two beliefs are observationally equivalent under a if they assign the same probability to each outcome that realizes with positive probability. This definition is equivalent to the one in the main text under Assumption 1(i).

The reason Theorem 1 only holds for myopic agents when we weaken (i) to (i′) is that Claim 1 can fail. The intuition is that even if the agent plays a many times, they may still think that playing a again will give them a non-trivial amount of information, as in the next example.

[30] When the agent is myopic Claim 1 continues to hold under Assumption 1(i′).

Example 8. *Let $A = \{a, b, c\}$, $Y = \{0, \bar{y}, y'\}$, and $\Theta = \{\bar{p}, p'\}$. Suppose that $\bar{p}_c(\bar{y}) = 1 - \bar{p}_c(0) = 0.9 = 1 - p'_c(0) = p'_c(y')$ and that $u(c, y) = -0.1$ for all $y \in Y$. Thus, the agent thinks that by playing c they pay a small cost, and with a very high probability they discover the correct model for sure, and otherwise receive an uninformative signal.*

For action b suppose that $\bar{p}_b(0) = 1 = p'_b(0)$ and $u(b, y) = 0$ for all $y \in Y$. That is, the agent thinks that action b is uninformative but safe.

Finally the agent thinks that action a produces the same information of action c but its payoffs are riskier: $\bar{p}_a(\bar{y}) = 1 - \bar{p}_a(0) = 0.9 = 1 - p'_a(0) = p'_a(y')$ $u(a, \bar{y}) = -100$ and $u(a, y') = 1$.

Here, c is not a a BN-E, because it is weakly dominated by action b, and it is never a myopic best reply. However, suppose that $p^*_c(0) = 1$, that the agent starts with a uniform prior over Θ, and the discount factor $\beta = \frac{1}{2}$. Then every optimal policy prescribes starting with action c to get information, and then switching to a forever after observing y', to b forever after observing $\bar{y}$ and trying c again after observing 0. Since $p^*_c(0) = 1$, the agent will continue to use c forever, because the believe that with high probability the true outcome distribution will be revealed next period.

Assumption 1(i) guarantees that when beliefs concentrate around a set of of outcome distributions that are observationally equivalent under a, i.e. $\nu \in \Delta(\mathcal{E}(a)(p))$ for some $p \in \Theta$, the experimentation value of a is weakly lower than that of some other action. This fact is used in Claim 1 to show that $G(\nu) > 0$ for every $\nu \in \Delta(\mathcal{E}(a)(p))$. Claim 1 holds under Assumption 1(i′) for myopic agents because for these agents all actions have 0 experimentation value.

Assumption 1(i′) is still sufficient for all the problems considered in Section 4.2. More generally, (i′) is sufficient when paired with with this additional assumption:

Assumption 2. $p, p' \in \mathcal{E}(a)(p) \Rightarrow p_a(y) = p'_a(y)$ for all $y \in Y$.

This assumption is trivially satisfied if all beliefs in the support of the agent's subjective prior assign positive probability only to signals which objectively occur with positive probability, i.e. $p_a(y) > 0 \Rightarrow p^*_a(y) > 0$ for all $p \in \Theta, a \in A$.

B.2 Extensions to Signals

Here we expand the probability space of our basic model in the obvious way: The sample space $\Omega = S^\infty \times (Y^\infty)^A$ 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. 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 $y_t = x_{a_t,s_t,k}$, where k is the number of times the agent has taken action a_t after signal s_t up and including period t. A (pure) policy $\pi : \bigcup_{t=0}^{\infty} S^{t+1} \times A^t \times Y^t \to A$ 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. For every $p, p' \in \Theta \cup \{p^*\}$, denote the supnorm distance between p and p':

$$||p - p''|| = \max_{s\in S, a\in A, y\in Y} |p_{a,s}(y) - p'_{a,s}(y)|.$$

Given our finite dimensionality assumption, the maximand depends on s only through the finite partition Ξ, so the supremum is attained. In this setting, a policy π converges to a strategy σ if there exists a T such that for all $t \geqslant T$, $\xi \in \Xi$, $p \in \Theta \cup \{p^*\}$ and $y \in Y$

$$\sum_{a\in A} \zeta\left(\left\{s \in \xi : \pi\left(a^T, y^T, s\right) = a\right\}\right) p_{a,s}(y) = \sum_{a\in A} \zeta\left(\{s \in \xi : \sigma(s) = a\}\right) p_{a,s}(y)$$

that is, there is finite time convergence over the behavior in the finite dimensional partition of signals considered by the agent. This restriction is without loss of generality if S is finite.

Lemma 10. *For every $\sigma \in A^S$ and $\varepsilon > 0$, $\hat{\Theta}(\sigma)$ and $\hat{\Theta}^{\varepsilon}(\sigma)$ are compact.*

Proof of Lemma 10. Compactness of $\hat{\Theta}(\sigma)$ follows from the generalization of Weierstrass Theorem to lower-semicontinuous functions (see e.g. Theorem 2.43 in Aliprantis and Border, 2013). Since the projection map is continuous it follows that $\hat{\Theta}^{\varepsilon}(\sigma)$ is closed, so it is compact. ■

Now we extend Lemma 2 to the case where the agent observes signals and has finite-dimensional beliefs. Since we restricted the policy function of the agent to be measurable in their beliefs, the set of policy functions is

$$\Pi = \left(A^S\right)^{\bigcup_{t=0}^{\infty}\left(A^t \times Y^t \times \Xi^t\right)}.$$

We endow the set A^S of measurable maps from S to A with the metric

$$d_\zeta\left(\sigma, \sigma'\right) = \zeta\left(\left\{s \in S : \sigma\left(s\right) \neq \sigma'\left(s\right)\right\}\right).$$

Then Π is the (countable) product space of measurable maps with index set $\bigcup_{t=0}^{\infty}\left(A^t \times Y^t \times \Xi^t\right)$.

Lemma 11. *Π is compact in the product topology, and for every $\nu \in \Delta(\Theta)$, $V\left(\cdot, \nu\right)$ is continuous with respect to the product topology.*

Proof. By Tychonoff's theorem A^S is compact in the product topology. Suppose that σ_n converges pointwise to σ, and let $C_n = \{s \in S : \forall m \geqslant n, \sigma_m\left(s\right) = \sigma\left(s\right)\}$. We have that $C_n \uparrow S$,

$$d_\zeta(\sigma_n, \sigma) = \zeta\left(\left\{s \in S : \sigma_n\left(s\right) \neq \sigma\left(s\right)\right\}\right) \leqslant 1 - \zeta(C_n)$$

and so $d_\zeta(\sigma_n, \sigma) \to 0$. Thus the product topology is finer than the topology induced by d_ζ, and so A^S is also compact in (A^S, d_ζ). Applying Tychonoff's theorem again, Π is compact in the product topology. Continuity follows from the fact that for every period $t \in \mathbb{N}$ the set $(A^t \times Y^t \times \Xi^t)$ is finite, and discounting. ■

We next generalize a couple of definitions given in the text to allow for signals. For every strategy σ and action contingent outcome distribution p, we let

$$p_\sigma = \int_S p^*_{\sigma(s),s}(\cdot) d\zeta(s))$$

denote the distribution over outcomes induced by the use of strategy σ. Let $\hat{\Theta}^\varepsilon(\sigma)$ denote the conceivable outcome distributions that are ε close to one of the elements of $\Theta(a)$:

$$\hat{\Theta}^\varepsilon(\sigma) = \{p \in \Theta : \exists p' \in \hat{\Theta}(\sigma), ||p'_\sigma - p_\sigma|| \leqslant \varepsilon\}.$$

Similarly, we denote the set of beliefs over conceivable distributions that assign at least probability $1-\varepsilon$ to $\hat{\Theta}^\varepsilon(\sigma)$ by

$$M_{\varepsilon,a} = \{\nu \in \Delta(\Theta) \colon \nu(\hat{\Theta}^\varepsilon(\sigma)) \geqslant 1 - \varepsilon\}.$$

Next we extend Lemma 3 to this setting.

Lemma 12. *If σ is a uniformly strict BN-E, then for every optimal policy π and every λ*

there exists an $\hat{\varepsilon} > 0$ such that for all $\varepsilon < \hat{\varepsilon}$

$$\nu \in M_{\varepsilon,\sigma} \implies |\zeta\left(\{s \in S : \pi\left(\nu, s\right) = a\}\right) - \zeta\left(\{s \in S : \sigma\left(s\right) = a\}\right)| < \lambda. \tag{5}$$

Proof. Fix a belief $\nu \in M_{\varepsilon,\sigma}$. Let π^{σ} denote the policy that always plays σ, and let Π_{λ} denote the set of policy functions $\tilde{\pi}$ such that:

$$|\zeta\left(\{s \in S : \tilde{\pi}\left(\nu, s\right) = a\}\right) - \zeta\left(\{s \in S : \sigma\left(s\right) = a\}\right)| \geqslant \lambda$$

Define $G(\varepsilon)$ as the gain from playing σ forever instead of using (one of) the best policies $\tilde{\pi} \in \Pi_{\lambda}$

$$G(\varepsilon) = \min_{\tilde{\pi} \in \Pi_{\lambda}} \min_{\nu \in M_{\varepsilon,a}} \left(V\left(\pi^{a}, \nu\right) - V\left(\tilde{\pi}, \nu\right)\right).$$

Notice that by Lemma 11 the space of the policy functions endowed with the product topology is compact. Since the subset of policy functions that satisfy 5 is closed, this subset is compact as well. Moreover, given that $\beta \in (0,1)$, the value function is continuous at infinity, and therefore $V\left(\pi^{a}, \nu\right) - V\left(\cdot, \nu\right)$ is a continuous function of the policy. Notice also that since $\mathbb{E}_{p,\pi}\left[\sum_{t=1}^{\infty}\left[\beta^{t-1}u(a_t, y_t)\right]\right]$ is continuous in p, $V\left(\pi^{a}, \cdot\right) - V\left(\tilde{\pi}, \cdot\right)$ is continuous in ν, so since $\varepsilon \to M_{\varepsilon,\sigma}$ is an upper hemicontinuous and compact valued correspondence, from the Maximum Theorem G is continuous in ε. Since σ is a uniformly strict BN-E, $G(0) > 0$, and there is an $\hat{\varepsilon}$ such that if $\varepsilon \leqslant \hat{\varepsilon}$, $G\left(\varepsilon\right) > 0$. This implies that for any optimal policy π it must be such that $\nu \in M_{\varepsilon,\sigma}$ implies that π satisfies (5), which proves the lemma. ■

Lemma 13. *Fix a strategy σ and $\varepsilon > 0$. There exists an $\bar{l} > 0$ such that for all $l \leqslant \bar{l}$ for every KL minimizer $q \in \hat{\Theta}(\sigma)$, every $p' \notin \hat{\Theta}^{\varepsilon}(\sigma)$, and every $\sigma' \in B_l(\sigma)$ we have*

$$f_l\left(\sigma', q, p'\right) := \sum_{y \in Y} p_{\sigma'}(y)\left(\frac{q_{\sigma'}(y)}{p'_{\sigma'}(y)}\right)^{l} > 1\,.$$

Proof. As noted by FII in their Lemma 3, for each KL minimizer $q \in \hat{\Theta}(\sigma)$ and every outcome distribution $p' \notin \hat{\Theta}(\sigma)$ there exists an $l\left(\sigma, q, p'\right)$ such that $f_l(\sigma, q, p') > 1$ for all $l \leqslant l\left(\sigma, q, p'\right)$. They also pointed out that for all $q, q' \in \Theta$, and $\sigma' \in A^{S}$, if $\hat{l} > l$ and $f_l(\sigma', q, q') \leqslant 1$, then $f_{\hat{l}}(\sigma', q, q') \leqslant 1$. We will now prove that there exists a uniform l that works for every $q \in \hat{\Theta}(\sigma)$ and $p' \in \hat{\Theta}^{\varepsilon}(\sigma)$, and every strategy σ' sufficiently close to σ.

Suppose by way of contradiction that there was no $\bar{l} > 0$ such that for all $l \leqslant \bar{l}$, $f_l(\sigma', q, p') > 1$ for all $q \in \hat{\Theta}(\sigma)$ and $p' \notin \hat{\Theta}^{\varepsilon}(\sigma)$, $\sigma' \in B_l(\sigma)$. Then we can define a sequence

(σ_n, q_n, p'_n) such that $f_{\frac{1}{n}}(\sigma_n, q_n, p'_n) \leqslant 1$, and $\sigma_n \in B_{1/n}(\sigma)$. The sequential compactness of $A^S \times \hat{\Theta}(\sigma) \times \overline{\{p \in \Delta(\Theta) \colon p_a \notin \hat{\Theta}^{\varepsilon}(\sigma))\}}$ derived in Lemma 10 guarantees that this sequence has an accumulation point (σ, q, p'). However, for, $n > \frac{1}{l(\bar{p}, p')}$, $f_{\frac{1}{n}}(\sigma_n, q_n, p'_n) \leqslant 1$ implies $f_{l(q,p')}(\sigma_n, q_n, p'_n) \leqslant 1$, but then the lower semicontinuity of $f_{l(q,p')}$ at (σ, q, p') leads to a contradiction with $f_{l(q,p')}(\sigma, q, p') > 1$. ■

Lemma 14. *Let $p, p', p^* \in \Delta(Y)$, and $l \in (0,1)$ be such that*

$$\sum_{y \in Y} p^*(y) \left(\frac{p(y)}{p'(y)} \right)^l > 1. \tag{6}$$

Then there is $\varepsilon' > 0$ such that for all $\nu \in \Delta(\Delta(Y))$, if we let

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

then

$$\sum_{y \in Y} r(y) \left[\left(\frac{\nu(B_{\varepsilon'}(p) \mid y)}{\nu(B_{\varepsilon'}(p') \mid y)} \right)^l \right] \geqslant \left(\frac{\nu(B_{\varepsilon'}(p))}{\nu(B_{\varepsilon'}(p'))} \right)^l.$$

for all $r \in B_{\epsilon'}(p^*)$

Proof. The lemma is trivially true if $\nu(B_\varepsilon(p')) = 0$ for some ε. Therefore, without loss of generality, we can assume that $\nu(B_\varepsilon(p')) > 0$ for all ε. Let $C_\varepsilon = B_\varepsilon(p^*) \times \Delta(B_\varepsilon(p)) \times \Delta(B_\varepsilon(p'))$ and define $G : \mathbb{R}_+ \to \mathbb{R}$ by

$$G(\varepsilon) = \min_{(r, \bar{\nu}, \nu') \in C_\varepsilon} \sum_{y \in Y} r(y) \left(\frac{\int_{B_\varepsilon(p)} \bar{q}(y) d\bar{\nu}(\bar{q})}{\int_{B_\varepsilon(p')} q(y) d\nu'(q)} \right)^l.$$

By the Maximum Theorem, the compactness of $\Delta(B_\varepsilon(p'))$ and $\Delta(B_\varepsilon(p))$ (see, e.g, Theorem 6.4 in Parthasarathy, 2005) and the fact that $G(0) > 1$ by equation (6), there is $\varepsilon' > 0$ such that for all $r, \nu', \bar{\nu} \in C_{\varepsilon'}$

$$\sum_{y \in Y} r(y) \left(\frac{\int_{B_{\varepsilon'}(p)} \bar{q}(y) d\bar{\nu}(\bar{q})}{\int_{B_{\varepsilon'}(p')} q(y) d\nu'(q)} \right)^l \geqslant 1. \tag{7}$$

Then,

$$\begin{aligned}
\sum_{y \in Y} r(y) \left(\frac{\nu(B_{\varepsilon'}(p) \mid y)}{\nu(B_{\varepsilon'}(p') \mid y)} \right)^l &= \sum_{y \in Y} r(y) \left(\frac{\int_{B_{\varepsilon'}(p)} \nu(B_{\varepsilon'}(p)) \bar{q}(y) d \frac{\nu(\bar{q})}{\nu(B_{\varepsilon'}(p))}}{\int_{B_{\varepsilon'}(p')} \nu\left(B_{\varepsilon'}(p')\right) q(y) d \frac{\nu(q)}{\nu(B_{\varepsilon'}(p'))}} \right)^l \\
&= \sum_{y \in Y} r(y) \left(\frac{\int_{B_{\varepsilon'}(p)} \bar{q}(y) d \frac{\nu(\bar{q})}{\nu(B_{\varepsilon'}(p))}}{\int_{B_{\varepsilon'}(p')} q(y) d \frac{\nu(q)}{\nu(B_{\varepsilon'}(p'))}} \right)^l \left(\frac{\nu(B_{\varepsilon'}(p))}{\nu\left(B_{\varepsilon'}(p')\right)} \right)^l \\
&\geqslant \left(\frac{\nu(B_{\varepsilon'}(p))}{\nu\left(B_{\varepsilon'}(p')\right)} \right)^l
\end{aligned}$$

where the inequality follows from equation (7). ∎

Theorem 1′. *Suppose the agent's beliefs are finite dimensional. Then if the strategy prescribed by the policy converges to σ with positive probability, then σ is a uniform BN-E.*

Proof. If σ is not a uniform BN-E, there is $\bar{p} \in \hat{\Theta}(\sigma)$ such that if $\operatorname{supp} \nu \subseteq \mathcal{E}_\sigma(\bar{p})$, then σ is not a myopic best reply to ν. We fix such a $\bar{p}$ throughout this proof.

Claim 5. *There exists $\varepsilon > 0$ such that if*

$$\frac{\nu\left(\{p \in \Theta \colon \forall s \in S, \forall y \in \operatorname{supp} p^*_{\sigma(s),s}, |p_{\sigma(s),s}(y) - \bar{p}_{\sigma(s),s}(y)| < \varepsilon\}\right)}{1 - \nu\left(\{p \in \Theta \colon \forall s \in S, \forall y \in \operatorname{supp} p^*_{\sigma(s),s}, |p_{\sigma(s),s}(y) - \bar{p}_{\sigma(s),s}(y)| < \varepsilon\}\right)} > \frac{1-\varepsilon}{\varepsilon},$$

then σ is not a myopic best reply to ν.

Proof. Define

$$G(\nu) = \max_\pi V(\pi, \nu) - \max_{\tilde{\pi} \colon \tilde{\pi}(\nu) = \sigma(\cdot)} V(\tilde{\pi}, \nu).$$

From the definition of $\bar{p}$, if

$$\operatorname{supp} \nu \subseteq \{p \in \Theta \colon \forall s \in S, \forall y \in \operatorname{supp} p^*_{\sigma(s),s}, p_{\sigma(s),s}(y) = \bar{p}_{\sigma(s),s}(y)\},$$

then $G(\nu) > 0$. By Lemma 11 the space of policy functions is compact and the value function is continuous in the policy, so $V(\cdot, \nu) - V(\cdot, \nu)$ is a continuous function of the policy, and since $\mathbb{E}_{p,\pi}\left[\sum_{t=1}^{\infty}\left[\beta^{t-1} u(a_t, y_t)\right]\right]$ is continuous in p, $V(\pi, \cdot)$ is continuous in ν. Therefore, we can conclude by the Maximum Theorem that G is continuous.

Now suppose that in contradiction to the claim, for every n there exists a ν_n such that

$$\frac{\nu_n\left(\{p\in\Theta\colon \forall s\in S,\forall y\in\operatorname{supp}p^*_{\sigma(s),s},\,|p_{\sigma(s),s}(y)-\bar{p}_{\sigma(s),s}(y)|<1/n\}\right)}{1-\nu_n\left(\{p\in\Theta\colon \forall s\in S,\forall y\in\operatorname{supp}p^*_{\sigma(s),s},\,|p_{\sigma(s),s}(y)-\bar{p}_{\sigma(s),s}(y)|<1/n\}\right)}\geqslant\frac{1-1/n}{1/n}$$

and $\sigma\in\pi\left(\nu_n\right)$. Because $\Delta\left(\Theta\right)$ is sequentially compact, $\left(\nu_n\right)_{n\in\mathbb{N}}$ has a converging subsequence $\left(\nu_{n_i}\right)_{i\in\mathbb{N}}\to\nu^*$. Thus, $\nu^*\left(\{p\in\Theta\colon \forall s\in S,\forall y\in\operatorname{supp}p^*_{\sigma(s),s},\,p_{\sigma(s),s}(y)=\bar{p}_{\sigma(s),s}(y)\}\right)=1$ and $G\left(\nu^*\right)=0$, which would imply that $\sigma\in\pi\left(\nu^*\right)$, a contradiction. ■

Now fix such an ε. Because the agent's beliefs are finite-dimensional, the agent believes that the outcome distribution depends on the signals only via the partition Ξ. We now define a finer partition of signals Ξ^σ such that for every two signals in the same cell i) the agent thinks they induce the same outcome distribution, i.e., they belong to the same cell of Ξ, and ii) σ prescribes the same action. Formally, Ξ^σ is the collection of subsets of signals of the form

$$\{s\in\xi_i\cap\sigma^{-1}\left(a\right)\text{ for some }\xi_i\in\Xi\text{ and }a\in A\}.$$

With a small abuse of notation, for every $\xi\in\Xi^\sigma$ let $\sigma\left(\xi\right)$ denote the action that strategy σ prescribes after every signal in ξ, and let $p_{a,\xi}$ be the probability distribution over outcomes induced under p after action a and any signal in ξ. Set $W=\Xi^\sigma\times Y$, and for each $p\in\Theta$, let p^σ be the unique probability measure over W that satisfies

$$p^\sigma\left(\xi,y\right)=\zeta\left(\xi\right)p_{\left(\sigma\left(\xi\right),\xi\right)}\left(y\right)\quad\forall\xi\in\Xi^\sigma,y\in Y.$$

Finally, define $\nu^\sigma\in\Delta\left(\Delta\left(W\right)\right)$ by

$$\nu^\sigma\left(C\right)=\nu\left(\{p:\bar{p}\in C\}\right)\quad\forall C\in\mathcal{B}\left(S\right)\times 2^Y.$$

For every $\eta\in(0,1)$, let

$$f_{\eta,q}=(1-\eta)p^{*\sigma}+\eta\bar{p}^\sigma.$$

Linearity of H in its first argument implies that for every $\eta\in(0,1)$,

$$p\in\operatorname*{argmin}_{p\in\Theta}H(f_{\eta,q},p^\sigma)\implies p^\sigma=\bar{p}^\sigma.$$

Let g be defined as in the main text with W replacing Y. We have

$$\begin{aligned}
& 2g\left((1-\eta)p^{*\sigma}+\eta\bar{p}^{\sigma},\varepsilon\right) \\
\geqslant\; & \inf_{q\in\Delta(W)\backslash B_{\varepsilon}(\bar{p}^{\sigma})}\sum_{w\in W}\left[(1-\eta)p^{*\sigma}(w)+\eta\bar{p}^{\sigma}(w)\right]\log q(w)-\sum_{w\in W}\left[(1-\eta)p^{*\sigma}(w)+\eta\bar{p}^{\sigma}(w)\right]\log\bar{p}^{\sigma}(w) \\
\geqslant\; & (1-\eta)\inf_{q\in\Delta(W)\backslash B_{\varepsilon}(\bar{p}^{\sigma})}\sum_{w\in W}p^{*\sigma}(w)\left[\log q(w)-\log\bar{p}^{\sigma}(w)\right] \\
& +\eta\inf_{q\in\Delta(W)\backslash B_{\varepsilon}(\bar{p}^{\sigma})}\sum_{w\in W}\bar{p}^{\sigma}(w)\left[\log q(w)-\log\bar{p}^{\sigma}(w)\right] \\
\geqslant\; & 0+\eta\inf_{q\in\Delta(W)\backslash B_{\varepsilon}(\bar{p}^{\sigma})}\sum_{w\in W}\bar{p}^{\sigma}(w)\left[\log q(w)-\log\bar{p}^{\sigma}(w)\right]\geqslant 2\eta\left(\varepsilon\right)^{2},
\end{aligned}$$

where the first inequality follows from the definition of g and the fact that the RHS minimizes over a larger set, the second inequality follows from concavity of the minimum, the third from the fact that $\bar{p}$ is a KL minimizer, and the fourth from Corollary 3.5 and Proposition 4.7 in Diaconis and Freedman (1990).

For every $t\in\mathbb{N}$, let $\eta_t=2t^{-\frac{1}{2}}$. If the empirical frequency is $f_{\eta_t,q}$ after t periods, and only strategy σ has been used, then from Lemma 8 and part (ii) of Assumption , there exists $\bar{g}>0$

$$\begin{aligned}
& \frac{\mu_t\left(\{p\in\Theta\colon\forall s\in S,\forall y\in\operatorname{supp}p^{*}_{\sigma(s),s},|p_{\sigma(s),s}(y)-\bar{p}_{\sigma(s),s}(y)|<\varepsilon\}\right)}{1-\mu_t\left(\{p\in\Theta\colon\forall s\in S,\forall y\in\operatorname{supp}p^{*}_{\sigma(s),s},|p_{\sigma(s),s}(y)-\bar{p}_{\sigma(s),s}(y)|<\varepsilon\}\right)} \\
=\; & \frac{\bar{\mu}_t\left(\{p\in\Theta\colon\forall w\in\operatorname{supp}p^{*\sigma},|p^{*\sigma}(w)-\bar{p}^{\sigma}(w)|<\varepsilon\}\right)}{1-\bar{\mu}_t\left(\{p\in\Theta\colon\forall w\in\operatorname{supp}p^{*\sigma},|p^{*\sigma}(w)-\bar{p}^{\sigma}(w)|<\varepsilon\}\right)} \\
\geqslant\; & \mu_0\left(\{p\in\Theta\colon\forall w\in\operatorname{supp}p^{*\sigma},|p^{*\sigma}(w)-\bar{p}^{\sigma}(w)|<\varepsilon^2\frac{2}{\bar{g}t^{\frac{1}{2}}}\}\right)\exp\left(t\eta_t\varepsilon^2\right)\geqslant\Phi\left(\varepsilon^2\frac{2}{\bar{g}t^{\frac{1}{2}}}\right)\exp\left(t^{\frac{1}{2}}\varepsilon^2\right).
\end{aligned}$$

By Lemma 7 there exists a $\hat{K},K'>0$ such that if the empirical frequency is f_t after t periods and $||f_{\eta_t,q}-f_t||<||\bar{p}^{\sigma}-p^{*\sigma}||t^{-\frac{1}{2}}/K'$ then

$$\frac{\mu_t\left(\{p\in\Theta\colon\forall s\in S,\forall y\in\operatorname{supp}p^{*}_{\sigma(s),s},|p_{\sigma(s),s}(y)-\bar{p}_{\sigma(s),s}(y)|<\varepsilon\}\right)}{1-\mu_t\left(\{p\in\Theta\colon\forall s\in S,\forall y\in\operatorname{supp}p^{*}_{\sigma(s),s},|p_{\sigma(s),s}(y)-\bar{p}_{\sigma(s),s}(y)|<\varepsilon\}\right)}\geqslant\Psi\left(\hat{K}\varepsilon^2\frac{2}{\bar{g}t^{\frac{1}{2}}}\right)\exp\left(\hat{K}t^{\frac{1}{2}}\varepsilon^2\right).$$

Fix an outcome $w^0\in\operatorname{supp}p^{*\sigma}$, and let f_t be the empirical frequency of the other $|\operatorname{supp}p^{*\sigma}|-1$ outcomes in the support of $p^{*\sigma}$. Denote by $p^{*\sigma}t$ the true probabilities of the same $|\operatorname{supp}p^{*\sigma}|-1$ outcomes.

An argument that mimics the proof of Claim 2 shows that $f_t \cdot t - p^{*\sigma} t$ is a $|\operatorname{supp} p^{*\sigma}| - 1$ dimensional random walk with nonsingular covariance matrix $\Sigma_{w,w'}$ for the increments.

By the Central Limit Theorem $(f_t - p^{*\sigma})\sqrt{t}$ converges to a Normal random variable with mean 0 and covariance matrix $\Sigma_{w,w'}$. Let $F_t = B_{\frac{||\bar{p}^{\sigma} - p^{*\sigma}||/K'}{\sqrt{t}}}\left(p^{*\sigma} + \frac{1}{\sqrt{t}}\left(\bar{p}^{\sigma} - p^{*\sigma}\right)\right)$. We have that

$$\mathbb{P}\left[f_t \in F_t\right] = \mathbb{P}\left[\sqrt{t}(f_t - \bar{p}^{*}) \in B_{||\bar{p}^{\sigma} - p^{*\sigma}||/K'}\left(\bar{p}^{\sigma} - p^{*\sigma}\right)\right].$$

Taking the limit $t \to \infty$ yields that

$$\lim_{t\to\infty} \mathbb{P}\left[f_t \in F_t\right] = \mathbb{P}\left[\tilde{Z} \in B_{||\bar{p}^{\sigma} - p^{*\sigma}||/K'}\left(\bar{p}^{\sigma} - p^{*\sigma}\right)\right]$$

where $\tilde{Z}$ is a random variable that is Normally distributed with mean $\vec{0}$ and covariance matrix $\Sigma_{w,w'}$. Consequently, if we denote as E_t the event that $f_t \in F_t$, it follows that $\sum_{t=1}^{\infty} \mathbb{P}\left[E_t\right] = \infty$. Moreover,

$$\liminf_{t\to\infty} \frac{\sum_{s=1}^{t}\sum_{r=1}^{t} \mathbb{P}\left[E_s \text{ and } E_t\right]}{\left(\sum_{s=1}^{t} \mathbb{P}\left[E_s\right]\right)^2} = \liminf_{t\to\infty} \frac{\frac{1}{t^2}\sum_{s=1}^{t}\sum_{r=1}^{t} \mathbb{P}\left[E_s \text{ and } E_r\right]}{\left(\frac{1}{t}\sum_{t=1}^{\infty} \mathbb{P}\left[E_t\right]\right)^2} \leqslant \liminf_{t\to\infty} \frac{\frac{1}{t^2}\sum_{s=1}^{t}\sum_{r=1}^{t} \mathbb{P}\left[E_r\right]}{\left(\frac{1}{t}\sum_{s=1}^{t} \mathbb{P}\left[E_s\right]\right)^2}$$
$$= \liminf_{t\to\infty} \frac{\frac{1}{t}\sum_{r=1}^{t} \mathbb{P}\left[E_r\right]}{\left(\frac{1}{t}\sum_{s=1}^{t} \mathbb{P}\left[E_s\right]\right)^2} = \frac{1}{\lim_{t\to\infty} \mathbb{P}\left[E_t\right]} = \frac{1}{\mathbb{P}\left[\tilde{Z} \in B_{||\bar{p}^{\sigma} - p^{*\sigma}||/K'}\left(\bar{p}^{\sigma} - p^{*\sigma}\right)\right]}.$$

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

$$\mathbb{P}\left[\bigcap_{t=1}^{\infty}\bigcup_{s=t}^{\infty} E_s\right] \geqslant \mathbb{P}\left[\tilde{Z} \in B_{||\bar{p}^{\sigma} - p^{*\sigma}||/K'}\left(\bar{p}^{\sigma} - p^{*\sigma}\right)\right] > 0\,.$$

The event $\bigcap_{t=1}^{\infty}\bigcup_{s=t}^{\infty} E_s$ is invariant under finite permutations of the increments $\left(\mathbf{1}_{w_t = w^1}, ..., \mathbf{1}_{w_t = w^{|\operatorname{supp} p^{*\sigma}|-1}} - p^{*\sigma}\right)$ with different time indices, so the 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 positive it must equal one.

This implies that $f_t \in F_t$ infinitely often with probability 1. It follows that the agent will eventually want to take an action different from σ:

$$\mathbb{P}\left[a_t \neq \sigma\left(s_t\right) \text{ for some } t\right] = 1\,.$$

Thus the strategy can not converge to σ with positive probability. ■

Theorem 2′. *Suppose σ is a uniformly strict BN-E. Then there is a belief $\nu \in \Delta(\Theta)$ such that for every $\kappa \in (0,1)$ there exists an $\varepsilon' > 0$ such that starting from any prior belief in $B_{\varepsilon'}(\nu)$:*

$$\mathbb{P}_\pi\left[\lim_{t\to\infty}\frac{1}{t+1}\sum_{r=0}^{t}\mathbf{1}_{\pi(a^r,y^r,s^{r+1})=\sigma(s_{r+1})} \geqslant 1-\kappa\right] > 1-\kappa\,.$$

Proof. Consider a uniformly strict BN-E σ, an optimal policy π and $\kappa \in (0,1)$. By Lemma 12, for every $\lambda \in (0,1)$ there exists an ε such that if $\nu(\hat{\Theta}^\varepsilon(\sigma)) \geqslant 1-\varepsilon$, then

$$|\zeta(\{s \in S : \pi(\nu, s) = a\}) - \zeta(\{s \in S : \sigma(s) = a\})| < \lambda.$$

For every $l \in (0,1)$, define the function $f_{l,\sigma} : P \times P \to \bar{\mathbb{R}}$ is defined by

$$f_l(\sigma', \bar{p}, p') = \sum_{y\in Y} p^*_{\sigma'}(y)\left(\frac{\bar{p}_{\sigma'}(y)}{p'_{\sigma'}(y)}\right)^l.$$

By Lemma 13, since $\hat{\Theta}^\varepsilon(\sigma)$ is compact by Lemma 10, and since f_l is lower semicontinuous, there exists $\varepsilon' \in (0,\varepsilon)$ such that $\bar{p} \in \hat{\Theta}^{\varepsilon'}(\sigma)$ implies that $f_l(\sigma,\bar{p},p') > 1$ for all p' with $p' \notin \hat{\Theta}^\varepsilon(\sigma)$. Let $K = \left(\frac{\varepsilon}{1-\varepsilon}\right)^l$. Then

$$\left(\frac{1-\nu\left(\hat{\Theta}^\varepsilon(\sigma)\right)}{\nu\left(\hat{\Theta}^{\varepsilon'}(a)\right)}\right)^l < K \implies \frac{1-\nu\left(\hat{\Theta}^\varepsilon(\sigma)\right)}{\nu\left(\hat{\Theta}^\varepsilon(\sigma)\right)} < \frac{\varepsilon}{1-\varepsilon}$$
$$\implies \nu\left(\hat{\Theta}^\varepsilon(\sigma)\right) > 1-\varepsilon \implies \pi(\nu) = a.$$

By Lemma 10, $\hat{\Theta}^\varepsilon(\sigma)$ is compact, so it has a finite cover $\{p \in \Theta : ||q^i_a - p_a|| \leqslant \varepsilon\}_{i=1}^n$, where $q^i \in \hat{\Theta}^\varepsilon(\sigma)$.

Let $\bar{\varepsilon}$ be such that $\nu\left(\hat{\Theta}^{\bar{\varepsilon}}(\sigma)\right) > 1-\bar{\varepsilon}$ implies that

$$\left(\frac{1-\nu\left(\hat{\Theta}^\varepsilon(\sigma)\right)}{\nu\left(\hat{\Theta}^\varepsilon(\sigma)\right)}\right)^l < \frac{K(1-\kappa)}{n}.$$

Then if the agent starts with a belief ν_0 with $\nu_0(\hat{\Theta}(\sigma)) > \bar{\varepsilon}$, σ is the unique best reply ν_0'. Moreover, by Lemma 14, Dubins' upcrossing inequality, and the union bound, there is a probability $(1-\kappa)$ that the positive supermartingale

$$\left(\frac{1-\nu_t'\left(\hat{\Theta}^{\varepsilon}(\sigma)\right)}{\nu_t'\left(\hat{\Theta}^{\varepsilon}(\sigma)\right)}\right)^l$$

never rises above K, and with probabilty $(1-\kappa)$

$$\left|\zeta\left(\{s \in S : \pi\left(\mu_t', s\right) = a\}\right) - \zeta\left(\{s \in S : \sigma\left(s\right) = a\}\right)\right| \leqslant \lambda,$$

for all $t \in \mathbb{N}$. Then the statement follows from the Hewitt-Savage $0-1$ Law ((see, e.g., Theorem 8.4.6 in Dudley, 2018). ■

Theorem 4′. *If signals are finite and subjectively uninformative and outcomes are subjectively exogenous, then any uniformly strict BN-E σ is positively attractive.*

Proof. Under the assumptions of the theorem, $\Theta \subseteq \Delta\left(\Delta\left(Y\right)\right)$. Consider a uniformly strict BN-E σ. By an obvious extension of Lemma 1 to the case with signals, $\Delta\left(\hat{\Theta}(\sigma)\right)$ is compact. Similarly, since S is compact and σ is the unique optimal best reply strategy at the beliefs in $\Delta\left(\hat{\Theta}(\sigma)\right)$, Lemma 3 can be extended to guarantee that there exists $\varepsilon \geqslant 0$ such that if

$$\nu\left(\mathrm{cl}\left(Q_{\varepsilon}\left(\bar{p}_{\sigma}\right)\right)\right) \geqslant (1-\varepsilon)$$

then the myopic best reply to ν is σ. By the same argument of the proof of Theorem 2, there exists an $l \in (0,1)$ and $\varepsilon' \in (0,\hat{\varepsilon})$, such that if $p \in Q_{\varepsilon'}\left(\bar{p}_{\sigma}\right)$ and $p' \notin Q_{\hat{\varepsilon}}\left(\bar{p}_{\sigma}\right)$ then $f_l(p,p') \geqslant 1$.

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

$$Q_{\hat{\varepsilon}/2}\left(\hat{p}_t\right) \subseteq Q_{\hat{\varepsilon}}\left(\bar{p}_{\sigma}\right).$$

Therefore by Proposition 1, 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 $\hat{\varepsilon}$ Ball around $\bar{p}$. That is,

$$\frac{\mu_{t'}(Q_{\hat{\varepsilon}}\left(\bar{p}_{\sigma}\right))}{1-\mu_{t'}(Q_{\varepsilon'}\left(\bar{p}_{\sigma}\right))} \geqslant \frac{\mu_{t'}(Q_{\hat{\varepsilon}/2}\left(\bar{p}_{\sigma}\right))}{1-\mu_{t'}(Q_{\varepsilon'}\left(\bar{p}_{\sigma}\right))} > 2\frac{(1-\hat{\varepsilon})}{\hat{\varepsilon}}.$$

Notice that replicating the outcome realizations y^t sufficiently many time yields a sequence

$y^{t'}$ such that the empirical frequency $\hat{p}_{t'} = \hat{p}_t$ and $t' > T$. Since $\operatorname{supp} p^*_{a,s} = Y$ for all $(a, s) \in A \times S$, this sequence of outcomes has positive probability, and after it occurs the agent plays σ. By Lemma 4 and the law of iterated expectations, conditional on a being played $\left(\frac{1-\mu_{t'}(Q_{\varepsilon'}(\bar{p}_\sigma))}{\mu_{t'}(Q_{\hat{\varepsilon}}(\bar{p}_\sigma))}\right)^l$ is a positive supermartingale.

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

Corollary 4. *Let α be a strongly uniform mixed BN-E in a problem (A, Y, p^*, u, Θ). There is a sequence of strategies $(\sigma_n)_{n \in \mathbb{N}}$ such that each $\sigma_{1/n}$ is a uniformly stable BN-E of a $(1/n)$ perturbation of (A, Y, p^*, u, Θ) and*

$$\lim_{n \to \infty} \zeta(\{s : \sigma_n(s) = a\}) = \alpha(a) \quad \forall a \in A.$$

If (A, Y, p^*, u, Θ) is subjectively exogenous and p^* has full support, the $\sigma_{1/n}$ can be chosen to be also positively attractive.

Proof. Let α be a mixed BN-E in a problem (A, Y, p^*, u, Θ). For every $n \in \mathbb{N}$, let $S = \operatorname{supp} \alpha$, $\zeta(a) = \alpha(a)$, and $\tilde{u}(a, y, s) = u(a, y) + \frac{1}{n}\mathbf{1}_{a=s}$, and let $\tilde{p}^*, \tilde{\Theta}$ be as given in part (ii) and (iii) of the definition of a perturbed environment.

Consider the strategy $\sigma(a) = a$. We have that for every $p \in \Theta$

$$\sum_{s \in S} \zeta(s) H\left(\hat{p}^*_{\sigma(s),s}, \Phi(p)_{\sigma(s),s}\right) = \sum_{a \in A} \alpha\left(a\right) p^*_a\left(y\right) \log p_a\left(y\right)$$

by (ii) and (iii) of the definition of a perturbed problem. Therefore, $\hat{\Theta}(\sigma) = \Phi(\hat{\Theta}\left(\alpha\right))$. Fix a signal $s \in S$, and consider any action $a' \neq \sigma(s)$. Since α is a strongly uniform BN-E

$$\mathbb{E}_{p_{\sigma(s)}}\left[u(\sigma(s), y)\right] \geqslant \mathbb{E}_{p_{a'}}\left[u(a', y)\right] \qquad \forall p \in \Theta(\alpha)$$

and by definition of $\tilde{u}$

$$\mathbb{E}_{p_{\sigma(s)}}\left[\tilde{u}(\sigma(s), y, s)\right] \geqslant \mathbb{E}_{p_{a'}}\left[\tilde{u}(a', y, s)\right] + 1/n \qquad \forall p \in \Theta(\alpha)$$

proving that σ is a strictly uniform BN-E. By construction

$$\zeta(\{s : \sigma_n(s) = a\}) = \alpha(a) \quad \forall a \in A.$$

Then the result follows by Theorems $2'$ and $4'$. ■

B.3 Additional Examples

Example 9 (A uniformly strict BN-E that isn't positively attractive)**.** *In this example the prior has support $\{p^1, p^2, p^3\}$. Here $a = 3$ is the only BN-E 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 details are in the following table:*

a	$a = 1$			$a = 2$			$a = 3$			$H(p^*_a, \cdot)$			$A^m(\delta_{(\cdot)})$
y	1	2	3	1	2	3	1	2	3				
u	1	0	0	0	1	0	0	0	1	$a = 1$	$a = 2$	$a = 3$	
p^*	0.1	0.9	0	0.9	0.1	0	0.1	0.1	0.8				
p^1	0.5	0.3	0.2	0.5	0.3	0.2	0.5	0.3	0.2	1.15	$\underline{0.74}$	2.03	$a = 1$
p^2	0.3	0.5	0.2	0.3	0.5	0.2	0.3	0.5	0.2	$\underline{0.74}$	1.15	2.03	$a = 2$
p^3	0.1	0.1	0.8	0.1	0.1	0.8	0.1	0.1	0.8	2.3	2.3	$\underline{0.64}$	$a = 3$

Example 10 (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 number of people at the market that day $s \in \{s_h, s_l\}$, with $s_h > s_l$. The payoff function is $u(a, y, s) = sy - 1_{a=a_H}$. 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 objectively optimal to hire a shop assistant after s_L, and it is objectively optimal to hire an assistant after s_H. The following argument *shows that the only BN-E is that the shop assistant is never hired: If the agent followed the objectively 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, this observation supports the belief that the shop assistant is useless. Since the myopic best reply to* δ_p *is to never hire the shop assistant, by Theorem 1' this suboptimal action is the only possible limit action.*