Dimensionality and Disagreement:

Asymptotic belief divergence in response to common information

Isaac Loh[*] Gregory Phelan[†]

This version: January 22, 2019

Abstract

We provide a model of boundedly-rational, multidimensional learning and characterize when beliefs will converge to the truth. Agents maintain beliefs as marginal probabilities rather than joint probabilities, and agents' information is of lower dimension than the model. As a result, for some observations agents may face an identification problem affecting the role of data in inference. Beliefs converge to the truth when these observations are rare, but beliefs diverge when observations presenting an identification problem are frequent. Robustly, two agents with differing priors who observe identical, unambiguous information may disagree forever, with stronger disagreement the more information received.

Keywords: Heterogeneous beliefs, divergence, learning, Bayesian updating, bounded rationality, sparsity.

JEL classification: D10, D80, D84.

[*] Northwestern University. Email: IsaacLoh2015@u.northwestern.edu .

[†] Department of Economics, Williams College. Email: gp4@williams.edu.

We are grateful for feedback from Matthew Chao, Eddie Dekel, Yingni Guo, Asim Khwaja, Rachel Kranton, Bruce Sacerdote, Larry Samuelson, Arunava Sen, and anonymous referees.

1 Introduction

People disagree and sometimes in big, persistent ways: people disagree about which policies best achieve outcomes, there is substantial disagreement among professional forecasters and central bankers regarding outlooks for macroeconomic variables,[1] and many patterns in financial markets strongly suggest investor disagreement (Hong and Stein, 2007). Furthermore, disagreements sometimes grow as people see the same information and continue to disagree. There is ample empirical evidence that people sometimes interpret evidence differently, whether they are Bayesian updaters or otherwise. [2] And yet, in many cases disagreements do *not* grow, and people appear to perceive evidence in similar ways (Gerber and Green, 1999). In this paper, we provide a model of multidimensional learning that can explain why disagreements can persist or even grow as agents observe identical information. Specifically, an agent's beliefs may differ from the true probabilities even after observing arbitrarily large amounts of information, and what beliefs converge to may depend on an agent's initial beliefs. As a result, observing common information can lead to permanent divergence in beliefs given arbitrarily small initial disagreements.

Our analysis requires two ingredients. First, agents sometimes face an identification problem for doing inference. We consider a multidimensional learning problem in which the world is of higher dimensionality than the signals (or information) people observe. In the sense of Benoît and Dubra (2018), some signals are "equivocal," meaning that the same signal can rationally be used to update beliefs in different ways depending on people's priors. Second, agents update beliefs about *marginal* probabilities, not the full joint-distribution. Because of the high dimensionality of the problem, boundedly-rational agents reduce the dimensionality of their learning problem by maintaining beliefs over marginal distributions and reconstructing joint distributions from the marginals using a fixed correlation (namely, independence).

[1] Andrade et al. (2016) find that disagreements among FOMC members about projections for the Fed Funds Rate and other variables are even greater than the disagreements among forecasters in the Survey of Primary Dealers. The point is not that central bankers and professional forecasters disagree, but that there is disagreement even within category (in which presumably there are similar information sets and objectives). Furthermore, economists rarely switch from hawks to doves (Malmendier et al., 2017).

[2] For example, Hirshleifer and Teoh (2003) document how the presentation of accounting information affects its interpretation, and Malmendier et al. (2017) show that personal experiences of inflation strongly influence the hawkish or dovish leanings of central bankers, which is evidence that priors influence how FOMC members interpret the same information. See Benoît and Dubra (2018) for a thorough discussion of the literature.

Our modification is driven by the observation that in multidimensional environments, maintaining a full joint prior distribution may be cognitively taxing. Fully rational agents would use Bayes' rule to update the joint distribution over the full state space rather than just considering the marginals and multiplying appropriately to form a joint distribution. Hence, *beliefs* about states would rarely be independent even if the underlying states are. Indeed, developing and maintaining a prior over the whole joint distribution of states of the world may be regarded as extremely computationally expensive, as suggested by Kominers et al. (2016). Additionally, Enke and Zimmermann (2017) provide strong experimental evidence that many people effectively ignore the need to update correlation when updating beliefs, and this effect is driven by complexity in the environment rather than the computational skills of the agents.

These two ingredients interact in important ways and we discuss them in turn. First, equivocal observations may lead to greater disagreement because agents use their current beliefs to update marginal beliefs. How might this mechanism look in reality? Consider two examples:

(i) A liberal and a conservative are watching the news together under a conservative government. After several reports that the country is weakening, the liberal says, "The government is failing us." The conservative, responds, "No, more evidence of the media's liberal bias."

(ii) Two economists, a Keynesian and a Neoclassical, are discussing the results of a recent stimulus package. The new GDP results are sluggish. The Neoclassical says, "Goes to show that stimulus doesn't work." The Keynesian replies, "Oh no, goes to show that the economy is much worse than we thought" (perhaps later adding that the stimulus was poorly designed).

In these cases, there are a number of underlying factors that contribute to the observed signals, but the signals are of lower dimensionality than the world. For the liberal and the conservative, the politics of news reflect the state of the country *and* how credibly that is reported. For the economists, GDP is a function of the fundamental strength of the economy, the effectiveness of stimulus, and potentially how well designed that stimulus was. But the observations do not identify those underlying variables. In each case, the exact same signal is interpreted in completely different ways, but the observers are not rejecting the information, nor is the information ambiguous or unclear: they are simply using the evidence in different, rational ways. Each observer uses the information to make inference about different underlying variables. The Keynesian infers that weak GDP means something about the economy, while the Neoclassical infers that weak GDP means the fiscal multiplier is low.

As we formalize below, the previous examples illustrate how agents reason when updating *marginal*, not joint, probabilities. The news-watchers are (together) correct that the news reports suggest *either* that "*country weak+media truthful*" or "*country strong+media biased*." But agents updating *marginal* beliefs effectively consider changing only one "dimension" of their beliefs at a time. So the liberal, believing *"weak+truthful"*, would compare that belief to "one-dimensional perturbations"—namely, "*strong+truthful*" and "*weak+biased*." The liberal rightly concludes that these beliefs are bad explanations of the news reports and thus chooses to stick with and even reinforce the initial belief. But the conservative, starting with different priors, would reason in precisely the same way to reach the opposite conclusion! However, if the two people instead considered the joint distribution, they would recognize that their beliefs about the country and the media should be *correlated*: the reporting suggests "*weak+true*" or "*strong+biased*." If they recognized this correlation and focused on only these two states, then they would have a chance of reaching agreement.

Second, agents would be justified in maintaining only marginal beliefs if there were no identification problems (i.e., if the signal space was sufficiently rich). If agents received signals separately for each dimension, then they could consider beliefs about each of those dimensions separately. However, when agents face identification problems then inference requires use of a joint distribution, which must somehow be reconstructed from marginals. As we discuss in greater detail below, one can think of misspecified learning as using the wrong method (i.e., copula) to reconstruct the joint distribution from marginals. In general, correctly reconstructing joint probabilities requires updating the method (i.e., copula) in light of new information.

Whether disagreements can persist depends on the extent to which agents can use the same observations to draw inference about different variables. Since agents may update beliefs differently when observing equivocal observations, the likelihood that disagreements persist depends on the severity of the identification problem and differences in their initial priors. When signals sufficiently identify the model, beliefs converge to the truth. But when identification problems are sufficiently severe, with *many* observations updated beliefs may converge but not necessarily to the true values (i.e., divergence). When divergence occurs, beliefs are likely to converge to values confirming initial beliefs whether those beliefs are correct or not. Thus, agents with heterogeneous priors may have beliefs diverge in light of common information, and agents with common priors may have beliefs converge to something other than the truth.

The rest of the paper is structured as follows. The remainder of this section discusses the related literature. Section 2 presents the baseline binary, two-state model, with theoretical results to characterize the limiting properties of beliefs. Section 3 presents a general version of this simple model and shows that the same intuitions hold in a more general setting. Section 4 discusses the setup and our results. Section 5 concludes.

Related Literature

Several papers have considered how observing a small number of signals can increase belief disagreement. Benoît and Dubra (2018) show that an equivocal signal will lead to *population polarization*, which is more systematic than *pairwise polarization*. We use equivocal signals as a primary ingredient and show that polarization results may hold *asymptotically* when people update marginal beliefs and equivocal observations are sufficiently frequent. Also in this literature, Baliga et al. (2013) show that polarization can occur as an optimal response to ambiguity aversion, and Andreoni and Mylovanov (2012) consider polarization about optimal actions when agents receive two-dimensional information to form a one-dimensional opinion. Generally, disagreements increase for signals with intermediate values but not for extreme values, which are more informative of the underlying structure. Similar observations are made by Dixit and Weibull (2007) and Jern et al. (2014).

Several papers study disagreement in Bayesian or boundedly rational settings. In Acemoglu et al. (2016), agents face an identification problem regarding how to interpret signals because agents do not know the noisiness of signals. In their model agents believe they will learn the true state asymptotically but not that they will *agree* asymptotically (the likelihood ratios of their beliefs need not converge). In our model agents need not learn the truth asymptotically, and agreement may not follow even when learning does. Fryer et al. (2018) consider when agents may receive "ambiguous" signals, which are interpreted in light of current priors (as in our model) and "stored" as an unambiguous signal. In our model signals are perfectly clear, but equivocal signals may be interpreted as evidence for several states of the world depending on current prior beliefs (i.e., dimensionality provides an explanation for why some signals may appear ambiguous). Finally, Baumeister and Hamilton (2015) consider VAR estimation in which sign restrictions partially identify the model (set identification). They show that Bayesian inference may continue to depend on priors even on identified sets and with arbitrarily large samples.

A critical ingredient in our model is that agents store only a fraction of the information in the joint distribution (see for example Gabaix (2014)). Any joint distribution can be decomposed into marginal probabilities together with a linking function (i.e., a copula) determining how to reconstruct joint probabilities from marginals. However, to correctly track the information in a joint distribution requires updating the parameters of the copula following observations. Thus, misspecification in our paper is that agents use a fixed, or static, copula, whereas a fully specified learning process would allow for a flexible, time-varying copula. Within this marginals-copula structure, our model features a boundedly-rational application of Bayes' Rule with common, exogenously determined observations. Other papers provide theories of inconsistent learning with a behavioral assumption (i.e., confirmatory bias, overconfidence, etc.), social learning, or endogenous signals.[3]

Finally, models allowing for asymptotic disagreement where Bayes' rule is used classically require carefully selected prior beliefs (disjoint support): agents that put belief zero on the truth will never learn the truth (see Esponda and Pouzo (2016) for a recent economic treatment of Bayesian learning with a misspecified prior). In our model, divergence is driven by the severity of the identification problem, which is a property of the fundamental state (not the coincidence of initial priors), and our model allows us to say when agents are guaranteed to learn the truth regardless of their priors.

[3] See for example Rabin and Schrag (1999), Eyster and Rabin (2010), Eyster et al. (2014), Schwartzstein (2014), Sundaresan and Turban (2014), Heidhues et al. (2015), Ortoleva and Snowberg (2015), and Sethi and Yildiz (2016).

2 A Simple 2×2 Model

This section provides a binary, 2-state model and characterizes asymptotic properties of boundedly rational multidimensional learning. This section illustrates the essential intuitions for our results in the general N-dimensional discrete model, which is presented in Section 3.

2.1 Setup

Time is indexed by $n = 1, 2, ...$ We suppress time subscripts when doing so does not create confusion.

Environment In each period, there are two random variables t_n and s_n that can each take values in $\{0,1\}$ (e.g., "failure" or "success" for each variable). The random variables are independently distributed, with distributions parameterized by two underlying state variables θ and σ: θ determines the frequency of success p_θ for t_n, σ determines the frequency of success q_σ for s_n, and realizations are independent across time. The state variables θ and σ also take binary values in $\{H, L\}$ (think "high" or "low" success rates). In particular, t-successes ($t = 1$) occur with higher probability when $\theta = H$, and s-successes occur with higher probability when $\sigma = H$. We denote the higher probabilities by p_H and q_H, and the lower probabilities by p_L and q_L. Thus,

$$\Pr(t = 1|\theta = H) = p_H > p_L = \Pr(t = 1|\theta = L),$$
$$\Pr(s = 1|\sigma = H) = q_H > q_L = \Pr(s = 1|\sigma = L).$$

All probabilities p_H, p_L, q_H, q_L lie strictly between zero and one.

The random variables t and s jointly determine a signal $Y_n = y(t_n, s_n)$. Specifically,

$$y(t,s) = t + s.$$

In period n an agent observes Y_n only but not the values of t_n and s_n. Clearly, a realization of $Y = 2$ or $Y = 0$ reveals the values of the random variables (i.e., both 1 or both 0). But a signal of $Y = 1$ is "equivocal" in the sense of Benoît and Dubra (2018): such an observation yields an identification problem since there are two possible combinations of random variables t and s that would provide that signal.

Hence, the state variables θ and σ determine the frequency of observations $Y_n \in \{0,1,2\}$. Comparing relative frequencies across states: when $(\theta,\sigma) = (H,H)$, then observations are relatively more likely to be twos and less likely to be zeroes; when (L,L), zeroes are relatively likely and twos unlikely; when (L,H) or (H,L), ones are comparatively likely, and these two states may differ from each other.

Beliefs An agent holds initial beliefs $P_0 = \Pr(\theta = H)$ and $Q_0 = \Pr(\sigma = H)$ and initial beliefs about θ and σ are independent. Agents may disagree about the likelihood of θ and σ, but not about how those states translate into realizations of t, s, or Y. In other words, the mapping from θ and σ to t and s is common knowledge.

Agents are boundedly rational in the following way. Rather than keeping track of the joint distribution of (θ,σ), which is a 2×2 matrix, agents store only the marginal probabilities P and Q. Given marginal beliefs, agents then reconstruct a joint distribution by assuming that the *beliefs* about the two states are *independent*. Hence, an agent believes the probability of $(\theta,\sigma) = (H,H)$ is PQ and the probability of $(\theta,\sigma) = (H,L)$ is $P(1-Q)$.

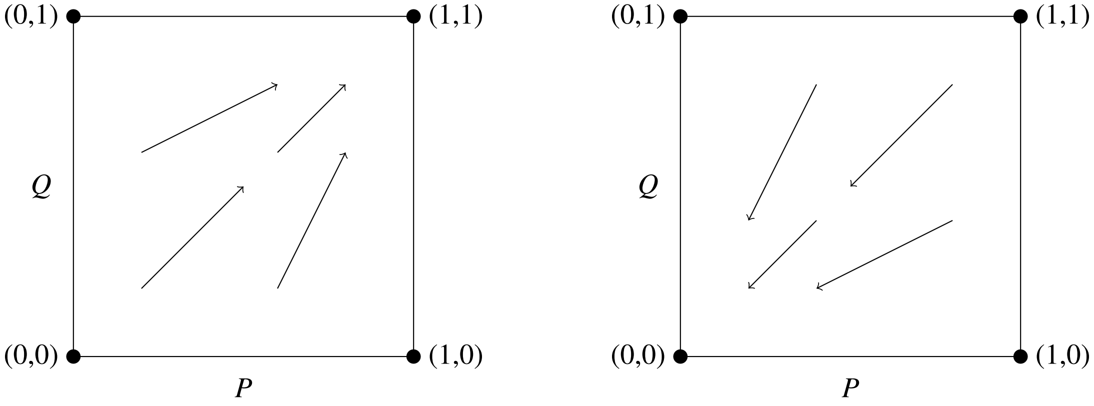
Note that keeping track of the full joint distribution requires three numbers: either the probabilities of three of the (θ,σ) states, or the two marginal probabilities together with the correlation coefficient to determine how marginal probabilities determine joint probabilities. Our boundedly rational agents keep track of only two values (the marginals) and neglect the correlation coefficient (keep it fixed). As we will discuss, even though the values of the aggregate states θ and σ are independent, that need not mean that an agent's *beliefs* about the states remain independent after receiving observations.

Belief Updating Storing beliefs in this way, agents use Bayes' Rule to *sequentially* update their marginal beliefs about the two states variables, using only the current observation and the current prior but not the complete history of observations. Thus, P_n and Q_n depend on Y_n, P_{n-1}, and Q_{n-1}, with joint probabilities determined by multiplying marginals appropriately.

It is convenient to derive the evolution of beliefs using odds ratios $O^P = \frac{P}{1-P}$ and $O^Q = \frac{Q}{1-Q}$. Let $O^P(Y)$ and $O^Q(Y)$ denote the updated odds ratios after observing Y. For observations $Y = 2$ and $Y = 0$, Bayes' Theorem applied to the marginal probability P yields

$$O^P_{n+1}(2) = \frac{p_H}{p_L} O^P_n \quad \text{and} \quad O^P_{n+1}(0) = \frac{1-p_H}{1-p_L} O^P_n,$$

and symmetric (with q_H and q_L) for Q. Hence, for these observations, agents update beliefs using the same likelihood function, becoming more optimistic/pessimistic about the aggregate states at the same time. Figure 1 (not drawn to scale) shows how beliefs move together in these cases. The figures plot beliefs in (P, Q) space with arrows indicating the (rough) directions that agents update beliefs. All agents update posteriors toward the same target, as standard learning models suggest, though posteriors will still differ because priors differ.



Updating marginals after $Y = 2$.

Updating marginals after $Y = 0$.

Figure 1: Updating marginal beliefs after $Y = 2$ or $Y = 0$. All agents become more optimistic/pessimistic about both states θ and σ, regardless of initial beliefs, jointly moving toward (H, H) for $Y = 2$ and to (L, L) for $Y = 0$.

However for the equivocal signal $Y = 1$, Bayes' Theorem yields

$$O^P_{n+1}(1) = \left(\frac{p_H(1-\overline{q}_n)+(1-p_H)\overline{q}_n}{p_L(1-\overline{q}_n)+(1-p_L)\overline{q}_n}\right)O^P_n, \quad O^Q_{n+1}(1) = \left(\frac{q_H(1-\overline{p}_n)+(1-q_H)\overline{p}_n}{q_L(1-\overline{p}_n)+(1-q_L)\overline{p}_n}\right)O^Q_n,$$

where (subscripts suppressed) $\overline{p} = Pp_H + (1-P)p_L$ and $\overline{q} = Qq_H + (1-Q)q_L$ are the ex-ante expected realizations of t and s in a period, given beliefs P and Q and assuming that beliefs about the aggregate states θ and σ are independent. Crucially, the likelihood ratios in this case depend on beliefs P and Q, and as a result equivocal realizations can lead to divergent posteriors (i.e., beliefs moving farther apart) when initial beliefs are sufficiently divergent (see Benoît and Dubra, 2018). Figure 2 (not drawn to scale) illustrates how beliefs diverge after observing $Y = 1$ depending on initial beliefs. As an example, when $p_H = q_H$ and $p_L = q_L$, then for $P > Q$, after observing $Y = 1$ the agent will increase P and decrease Q, but the opposite will occur when $Q > P$, and when $P = Q$ beliefs will not change. More generally, if $p_H \neq q_H$ and $p_L \neq q_L$ then there exists a "divergence curve" such that beliefs "near" (L,H) update toward (L,H), and similar for beliefs near (H,L).

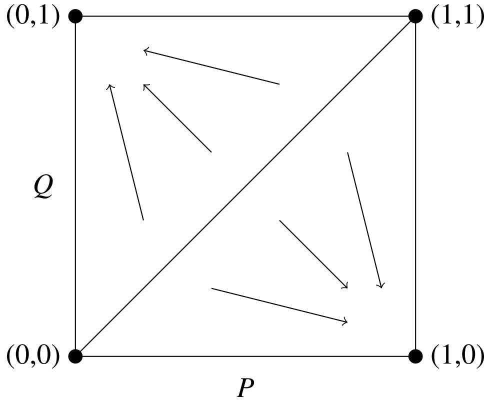


Figure 2: Updating marginal beliefs after $Y = 1$. Change in optimism/pessimism about each state depends on initial beliefs. Initial relative beliefs about P,Q get reinforced, moving either toward (H,L) or toward (L,H); $p_H = q_H$ and $p_L = q_L$.

Divergent updating in light of $Y = 1$, an example of the result in Benoît and Dubra (2018), should not be surprising. When $Y = 1$, agents cannot identify the underlying values of t and s but instead use their beliefs to infer which values of t and s are most likely. When $P = Q$ and states are symmetric, an agent believes $t = 1$ or $s = 1$ are equally likely, and so $Y = 1$ provides no information about the *relative* likelihood of θ and σ. Crucially, however, for a classical joint-updating Bayesian, $Y = 1$ *does* provide information about the *joint* distribution of (θ, σ). In particular, the states (L,H) and (H,L) would be relatively more likely than (H,H) and (L,L), since these states are relatively more likely to generate two (or zero) successes than just one.

Contrast with Classic Bayesian Updating Before proceeding to the asymptotic results, it is worth contrasting how learning works if agents maintain the full joint distribution and update using Bayes in the classical way (i.e., "joint-updating"). With joint-updating, beliefs converge to the truth whether agents update sequentially or using the full history. To simplify, suppose for now that the aggregate states θ and σ are nearly symmetric, so that $p_H \approx q_H$ and $p_L \approx q_L$. (The following reasoning carries over if the states are asymmetric, with some slight complications.)

Now consider how beliefs would rationally update after a string of observations. If many observations are $Y = 0$, then beliefs would update so that the most likely state is (L,L), the least likely state is (H,H), and both (L,H) and $(1,0)$ may remain somewhat likely (though with enough observations they would not be). Similarly, if many observations are $Y = 2$, then beliefs would update so that (H,H) would be very likely and (L,L) would be unlikely. Figure 3 illustrates the changes in beliefs of the joint distribution following these observations, where each box corresponds to the probability of the state in its unique corner (e.g., the top-right box is $\Pr(H,H)$ and the lower-right box is $\Pr(H,L)$). Notice that in both of these cases, beliefs about θ and σ would become *positively correlated* with joint-updating. Importantly, the belief that $\Pr(H,H)$ is higher and $\Pr(L,L)$ is lower can be approximated by letting P and Q each be higher (or lower). Indeed, this is how marginals-updating would work in light of these observations. Hence, the joint distributions in these cases can be nearly reconstructed using independent beliefs together with appropriate marginals.

In contrast, a string with many $Y = 1$ would make the intermediate states (L,H) and (H,L) most likely, because in these states success for one variable together with failure for the other is likely. Similarly, the beliefs about states (L,L) and (H,H) would be low. In this case, beliefs about θ and σ would be *negatively* correlated, as illustrated in Figure 4. However, marginals-updating would move beliefs strictly toward *either* (L,H) or (H,L), depending on beliefs. When $P > Q$,

$\sigma = H$	-	↑
$\sigma = L$	↓	-
$\Pr(\theta, \sigma)$	$\theta = L$	$\theta = H$

Updating beliefs after $Y = 2$.

$\sigma = H$	-	↓
$\sigma = L$	↑	-
$\Pr(\theta, \sigma)$	$\theta = L$	$\theta = H$

Updating beliefs after $Y = 0$.

Figure 3: Joint-updating beliefs become positively correlated after observing $Y = 2$ or $Y = 0$. Beliefs move in same direction as marginals-updating beliefs.

then beliefs would move toward (H, L), as illustrated in the figure.

$\sigma = H$	↑	↓
$\sigma = L$	↓	↑
$\Pr(\theta, \sigma)$	$\theta = L$	$\theta = H$

Joint-updating beliefs.

$\sigma = H$	↓	–
$\sigma = L$	–	↑
$\Pr(\theta, \sigma)$	$\theta = L$	$\theta = H$

Marginals-updating with $P > Q$.

Figure 4: Joint-updating beliefs become negatively correlated after observing $Y = 1$. Marginals-updating beliefs move strictly toward (H, L), reinforcing initial beliefs.

Crucially, with independent beliefs over θ and σ (zero correlation) *no changes in marginal probabilities are able to replicate this change in the joint distribution.* With independent beliefs, the probability of either (H, L) or of (L, H) can be high, but both cannot be high at the same time without the probabilities of (H, H) and (L, L) also being high. Thus, a marginals-updating agent is not able to focus on the combination of states which are observationally similar, namely when $\theta \neq \sigma$. If $p_H \neq q_H$ and $p_L \neq q_L$, then a rational, joint-updating Bayesian will "quickly" learn that $\theta \neq \sigma$, since 1's are prevalent. Once the joint-updating agent has narrowed the states down to these two, the agent will be able to learn the difference between them over time by observing any slight difference in the frequencies of Y arising from differences among p_H, p_L, q_H, and q_L. As we will discuss, a boundedly rational, marginals-updating Bayesian, who assumes a constant correlation of beliefs, does not update beliefs in this way.

This example illustrates a key difference between joint- and marginals-updating. Marginals-updating agents consider changes in only one dimension of beliefs at a time rather than jointly considering how to update P and Q together. Accordingly, beliefs are restricted in how they can update: correlation is fixed but should update, and so belief updates are effectively limited to "local perturbations." In contrast, joint-updating agents consider how the probability of each state ought to change, and so marginal probabilities change in a correlated manner.

2.2 Theoretical Results in the Simple Model

In this section we characterize when beliefs are guaranteed to be consistent (i.e., they converge to the truth), and when inconsistent learning is possible (i.e., beliefs need not converge to the truth).

2.2.1 Consistent Learning

First, agents' beliefs converge to the correct beliefs whenever $\theta = \sigma$.

Proposition 1. *If* $(\theta, \sigma) = (H, H)$ *or* (L, L)*, then (respectively)* $P_n, Q_n \overset{a.s.}{\to} 1$ *and* $P_n, Q_n \overset{a.s.}{\to} 0$.

In these cases, unequivocal observations ($Y = 2$ or $Y = 0$) are relatively most likely. The states (H, H) and (L, L) are not observationally similar to any other state. Thus, the (relatively) most frequent observations provide clear indication of the state. Since the likelihood functions for $Y = 0$ and $Y = 2$ do not depend on beliefs, and because these observations are relatively frequent, beliefs will drift toward the true state regardless of initial beliefs.

However, there is a second observation to be made that will be critical to the analysis in the general model: when $\theta = \sigma$, even "local" perturbations of a belief will dominate the prior. Consider many observations of $Y = 2$ and how an agent would "reason" about the likelihood of the different states. Suppose the agent first considered the state (L, L) and in light of the evidence asks, "Are any adjacent states more likely?" or in other words, "Can I change just one dimension and do better?" Compared to (L, L), both (H, L) and (L, H) appear more likely after many $Y = 2$ (since both are more likely to produce successes). Thus, a belief that increases only *one* of P or Q would better explain the data. From there the agent may ask the same question starting from (H, L) or (L, H), and would now be willing to move toward (H, H), since doing so explains the data still better.

Thus, even if belief updating were restricted to these "local perturbations," beliefs would nevertheless update toward the true state since these local changes move marginal beliefs in the right direction. Even though observations of $Y = 2$ or $Y = 0$ should induce *positive* correlation in beliefs, an agent will want to update marginals in the same (i.e., correlated) direction, and so marginals-updating agents will learn the truth even when the correlation of beliefs is fixed.

Even when learning is asymptotically consistent, that need not mean that agents learn at the same rate (this result is similar to the result in Acemoglu et al. (2016)). When $\theta = \sigma$, both P and Q converge to the truth but agents starting with different priors will continue to hold different *relative* beliefs about P and Q asymptotically—and those relative beliefs will diverge—even as marginal beliefs converge to the truth (see Lemma 8).

2.2.2 Inconsistent Learning

In contrast, when equivocal observations $Y = 1$ are relatively likely, as occurs when $\theta \neq \sigma$, beliefs needs not converge to the truth. Divergence may occur when the states (H,L) and (L,H) produce outcomes that are observationally *similar*, even if not identical. Thus, in cases when a classical, joint-updating Bayesian would be able to learn the true state, a boundedly rational, marginals-updating agent may not learn the truth. Individuals can converge asymptotically to false beliefs even when states are asymmetric.

Proposition 2. *There is a positive measure set of parameters such that for any priors* $0 \leq Q_0 < P_0 \leq 1$*, with positive probability beliefs do not converge to the true state when* $\theta \neq \sigma$*.*

Proposition 2 states that robustly there is no guarantee of marginals-updating converging to the truth. In particular, (i) there exist parameters such that *there exist* priors such that *almost surely* beliefs converge to the wrong values; (ii) there exist parameters such that *any* priors will converge to the wrong values with *positive* probability; (iii) there exist parameters such that *some* priors are guaranteed to converge to the wrong values with *positive* probability.

In light of the consistency result in Proposition 1, what is the intuition for why belief divergence occurs in this case? First, when $\theta \neq \sigma$ equivocal observations are relatively most frequent, and thus how agents update beliefs will frequently be determined by their current beliefs and (as discussed above) will reinforce those beliefs. Second, "local" perturbations of beliefs do not necessarily dominate a given prior. If an agent believes that (L,H) is likely, neither (L,L) or (H,H) need to appear as better explanations of the observed data, in which 1's are relatively frequent. Thus, an agent considering (L,H) and restricted to only update one belief dimension at a time would not choose to do so. Choosing (H,L) over (L,H) requires making two changes to marginal beliefs. However, if an agent were only considering the two states (L,H) and (H,L) (ruling out the others, as a classical, joint-updating Bayesian would), then an agent would be willing to update beliefs in the direction of the state that better explains the observed data (namely, the true state). But since agents update only marginals, maintaining fixed correlation of beliefs, a belief that (L,H) is likely may be locally dominant.Thus, marginals may update to make (L,H) more likely and (H,L) less likely.

These theoretical results are expanded in Appendix A, and in Appendix C we provide simulation evidence indicating that divergence is most likely to occur in a neighborhood of parameters around symmetry, when the states (H,L) and (L,H) are observationally more similar. Indeed, divergence is guaranteed when the states are exactly symmetric.

Corollary 1 (Asymptotic belief divergence)**.** *Let $p_H = q_H$ and $p_L = q_L$. Suppose $\theta \neq \sigma$. If $P_0 > Q_0$, then $P_n \overset{a.s.}{\to} 1$ and $Q_n \overset{a.s.}{\to} 0$ (vice versa if $P_0 < Q_0$). Furthermore, suppose one agent holds prior beliefs with $P_0 > Q_0$ and the other has priors with $P_0 < Q_0$. Then if $\theta \neq \sigma$, with probability 1 agents' beliefs will asymptotically diverge to complete polarization, the first with $P_n \overset{a.s.}{\to} 1$ and $Q_n \overset{a.s.}{\to} 0$, and the reverse for the other.*

In the perfectly symmetric case, a joint-updating Bayesian would learn that $\theta \neq \sigma$ (those joint probabilities would sum to 1), but beliefs would not converge toward one state or the other because they are observationally equivalent—the data are not rich enough to distinguish. In contrast, a marginals-updating agent *will* converge to certainty, with posterior beliefs confirming initial relative priors (i.e., complete polarization occurs asymptotically).

3 General Model

This section describes a general multivariate model with dimension $d \geq 2$. As in the simpler model, the model we describe here is almost identical to typical Bayesian inference with the only modification being that our agents do not retain a joint distribution over a multidimensional state space but store only marginals.

3.1 Setup

Let there be a multivariate state of the world $\theta = (\theta_1, \dots, \theta_d)$ taking values in a discrete set $\Theta = \prod_{i=1}^{d} \Theta_i$. Let $Y_n \in \mathbb{R}$ be the period n signal and F_θ be its distribution function when the state of the world is θ; we suppose that Y_n is drawn independently from F_θ in every period. Let f_θ denote the corresponding density of Y_n. We assume that a true state $\theta^* \in \Theta$ is randomly chosen ex-ante from a product probability distribution $\mathrm{P} = \mathrm{P}_1 \times \cdots \times \mathrm{P}_d$ on $\Theta = \Theta_1 \times \cdots \times \Theta_d$. Initial priors W_0^i are fixed and define a product probability distribution over Θ.

As before, in every period the agent updates the marginal probabilities for each θ_i and retains only those marginals, not the full joint distribution. Let $W_n^i(\hat{\theta}_i) \equiv W_n(\theta_i = \hat{\theta}_i)$ denote the agent's period-n belief that $\theta_i = \hat{\theta}_i$ and W_n^i the resulting marginal measure over Θ_i. These will be the objects which are updated in every period. Applying Bayes? theorem requires a joint distribution, so the agent attempts to construct the joint distribution using some choice of copula Ψ, which we define to be a function mapping d marginal prior distributions $W_n^1, \dots, W_n^d$ to a joint prior distribution:

$$W_n(\theta = (\hat{\theta}_1, \dots, \hat{\theta}_d)) = \Psi\big(W_n^1, \dots, W_n^d\big)(\hat{\theta}_1, \dots, \hat{\theta}_d). \tag{1}$$

We suppose the agent uses the independence copula to reconstruct a joint distribution; hence, $W_n \equiv \prod_{i=1}^{d} W_n^i$ is the induced product probability measure over Θ (i.e., the agent multiplies marginal probabilities). Let $W_{n+1}(\hat{\theta}_i) \equiv W_{n+1}(\hat{\theta}_i | Y_{n+1})$ denote the agent's updated prior upon observing signal Y_{n+1} with prior W_n. In place of standard probability notation, we use W_n to denote the agent's beliefs over marginals and as an induced probability measure on the space Θ. More precisely, for every n, W_n is a measure on Θ reflecting the agent's prior over Θ given the sequence of observations $Y_n, Y_{n-1}, \dots$. It is derived along Bayesian updating rules, but for the restriction that agents update marginal beliefs in every period using some function (i.e., copula), represented here by Ψ, to reconstruct the joint prior. We use standard probability notation for W, e.g.

$W_n(\theta_i = \hat{\theta}_i) \equiv W_n(\{\theta \in \Theta : \theta_i = \hat{\theta}_i\})$. Appendix B provides derivations as well as a motivation for storing marginal probabilities and using the independence copula as a result of scarce memory (Remark 1).

3.2 Theoretical Results in the General Model

In the general model, we now consider situations under which the learning problem enables the agent to learn the true state, and those in which learning fails. Statements of "almost surely" are to be interpreted as with respect to the resulting probability distribution over the space of possible signal sequences $\{Y_n\}_{n=1}^{\infty}$.

3.2.1 Consistent Learning

We have seen in the 2×2 model that agents would learn the true state whenever the state is either (L, L) or (H, H) (i.e., learning is asymptotically consistent). We discussed two characteristics of these states that made learning with marginals-updating possible: (i) the states were observationally distinct from other states (i.e., the identification problem was not severe because equivocal signals were relatively rare), and (ii) even local perturbations of beliefs would lead agents to prefer the true state regardless of what beliefs they start with. In this general setting, marginals-updating agents will learn the true state when these two conditions, formally defined, are met for that state.

In order for agents to learn a state θ^* thus requires the following conditions. First, we require that the identification problem not be severe for the state. Extreme values of Y must be strong evidence for the state θ^* in terms of the likelihood ratio. In the 2×2 model, observations of 2 were strong evidence in favor of the state (H, H), and observations of 0 in favor of the state (L, L).

A1: For the state $\theta^* = (\theta_1^*, \ldots, \theta_d^*)$ the likelihood ratio $\frac{f_{\theta^*}(y)}{f_\theta(y)}$ is non-decreasing in y for all $\theta \in \Theta$, and $W_0(\theta^*) = \prod_{i=1}^d W_0^i(\theta_i^*) > 0$.

Condition **A1** is the condition that the state is observationally distinct from other states (i.e., the identification problem is not severe). The condition implies that the state θ^* is easy to learn in the sense that higher (or lower) observations of Y provide a stronger signal that the true state is θ^*, and

θ^* generates a relatively high frequency of high (or low) observations of Y. The assumption also requires that the agent's initial prior places positive mass on θ^*.

Second, we require that for any belief, even local perturbations of beliefs will favor the state θ^* in the sense of being better explanations of the distribution of observations Y.

A2: For any $\theta = (\theta_1, \ldots, \theta_d)$, let $T_i(\theta) = (\theta_1, \ldots, \theta_i^*, \ldots, \theta_d)$ be θ with the i^{th} coordinate replaced by θ_i^*. Then $F_{T_i(\theta)}(y)$ first-order stochastically dominates $F_\theta(y)$ for all θ and i.

Condition **A2** states that switching a single dimension θ_i to its true value θ_i^* implies a better statistical explanation of the signals. In the 2×2 case, we said that when the true state is (H,H) then a local perturbation of the belief (L,L) would provide a better explanation of the observed data, as would local perturbations of (H,L) and (L,H). This condition is the multidimensional, generalized version of that insight.

Finally, to guarantee learning in these cases requires an innocuous but technical regularity condition bounding the distribution of signals, which we discuss further in Appendix B.2:

A3: $|\Theta| < \infty$. There is a set $A \subset \mathbb{R}$ satisfying $\Pr(y \in A | \theta = \theta^*) > 0$ such that for all $y \in A$, $i \in \{1, \ldots, d\}$, and $\theta \in \Theta$,

$$\frac{f_{T_i(\theta)}(y)}{f_\theta(y)} \geq 1 \quad \text{and} \quad \frac{f_{\theta^*}(y)}{f_\vartheta(y)} > 1 \quad \text{for all } \vartheta \neq \theta^*.$$

Given these conditions, agents will learn the true state θ^* asymptotically. States satisfying these conditions are sufficiently observationally different from other states so that updating marginal probabilities with a fixed copula is sufficient to distinguish these states from other states.

Theorem 1 (Consistency)**.** *Suppose that f_θ is continuously differentiable for every θ, agents update marginals W_n according to the independence cupola, and the true state is θ^*. If **A1** and **A2** hold then almost surely $\lim_{n\to\infty} W_n(\theta_i^*)$ exists and is strictly positive. If **A3** also holds then for every i, $W_n(\theta_i = \theta_i^*) \overset{a.s.}{\to} 1$.*

The conditions for consistent learning are more restrictive than the conditions for the inconsistency argument below (Theorem 2). This might indicate that when updating occurs along marginals, with a fixed copula, asymptotic inconsistency is more common than consistency of beliefs. However there is one situation in which the conditions arise naturally, which is when the aggregate signal Y is a sum of d random variables X_i for which the distribution of X_i under θ_i^* first-order stochastically dominates the distribution under other θ_i, for all $i = 1, \ldots, d$. The following is a simple subcase of Theorem 1 which arises when the distribution of Y is a convolution of the X_i distributions:

Corollary 2. *Let $X_i \sim G_{\theta_i}$ for every i where G_{θ_i} has continuously differentiable density g_{θ_i}. Suppose that $y = \sum_{i=1}^{d} X_i$ (with corresponding density $f_\theta = g_{\theta_1} * \cdots * g_{\theta_d}$) and for all i, $\theta_i \in \Theta_i$, the ratio $\frac{g_{\theta_i^*}(x_i)}{g_{\theta_i}(x_i)}$ is weakly increasing. Then $W_n(\theta_i^*)$ is a submartingale for every i which converges almost surely in* $(0,1]$.

This result is illustrated in the 2 × 2 case by Proposition 1.

3.2.2 Inconsistent Learning

Consistent learning required that the identification problem not be severe and that local perturbations of beliefs around any state would favor the truth. Inconsistent learning, thus, may arise if there is an identification problem and if local perturbations of beliefs around *some* states are not guaranteed to favor the true state.

First, we require that there is no observation of Y that would perfectly identify any state θ:

B1: Densities are relatively bounded:

$$\sup_{y,\theta,\theta'} \frac{f_\theta(y)}{f_{\theta'}(y)} < \infty. \tag{2}$$

Condition **B1** merely states that no aggregate state is perfectly identified by any observation Y. This is a very general restriction. The underlying random variables X may be perfectly revealed by Y, but that is not the same as perfectly identifying the underlying state θ. In the 2×2 model, no Y perfectly identified (θ, σ) even though $Y = 2$ and $Y = 0$ perfectly reveal the values of the random variables t and s. Importantly Condition **A1** and condition **B1** can both hold at the same time.

Second, recall that an agent's beliefs may converge to the wrong state $\hat{\theta}$ if local perturbations of a belief vector near $\hat{\theta}$ (moving only a single dimension of a belief at a time) do not provide a better explanation of the data. The true state θ^* would provide the best explanation of the data, but moving beliefs to θ^* from $\hat{\theta}$ would require changing multiple dimensions in the belief vector. This is the essence of Condition **B2**, which requires defining some objects before we can state it formally. Importantly, Condition **B2** will impose properties on the behavior of beliefs around $\hat{\theta}$ *when* the true state is θ^*.

Consider neighborhoods of a state $\hat{\theta} \in \Theta$ that may be obtained by changing one of its elements at a time (i.e., "local perturbations"). For every $i = 1, \ldots, d$, we set

$$G_i(\hat{\theta}) = \{\theta \in \Theta : \theta_{-i} = \hat{\theta}_{-i}, \theta_i \neq \theta_{-i}\},$$

i.e., for all $\theta \in G_i(\hat{\theta})$, $\theta_j = \hat{\theta}_j$ for every $j \neq i$. Given probability densities f, g on a set X with dominating measure μ, the *Kullback-Leibler* divergence of f and g is given by

$$D_{\mathrm{KL}}\left(f \,||g\,\right) = \int_X f(x) \log \frac{f(x)}{g(x)}\, \mathrm{d}\mu.$$

The *Kullback-Leibler* divergence (relative entropy) is a weighted sum of the log-likelihood ratios for two distributions, weighted by the frequency of observations, and is a standard measure of information loss.[4] Given these definitions, we can now formally define *local dominance*.

B2: The belief $\hat{\theta}$ is *locally dominant* for *i*:

$$D_{\mathrm{KL}}\left(f_{\theta^*} \,\middle|\middle|\, f_{\hat{\theta}}\,\right) < \inf_{f \in \mathrm{co}\left(\{f_\theta : \theta \in G_i(\hat{\theta})\}\right)} D_{\mathrm{KL}}\left(f_{\theta^*} \,||f\,\right). \tag{3}$$

where $\mathrm{co}\,(\cdot)$ denotes the convex hull.

Condition **B2** then says that when the true state is θ^*, the state $\hat{\theta}$ does a better job describing the distribution of X compared to every *local perturbation* of $\hat{\theta}$ (and including distributions in the convex hull of the perturbations). The the divergence of $f_{\hat{\theta}}$ from f_{θ^*} measures how much worse $\hat{\theta}$

[4] See Esponda and Pouzo (2016) for an application of *Kullback-Leibler* divergence in Berk-Nash equilibria.

does at describing the distribution of random variables X compared to the actual distribution. Then condition requires considering the probability distributions associated with all local perturbations of $\hat{\theta}$ by considering all states $\hat{\theta}'$ that agree with $\hat{\theta}$ at all but one coordinate.

Importantly, $\hat{\theta}$ need not be the best explanation of the world globally (indeed, θ^* is), but locally $\hat{\theta}$ is the best when considering changes in only one coordinate at a time ($\hat{\theta}$ makes more sense than local perturbations of $\hat{\theta}$). Thus, a state $\hat{\theta}$ is locally dominant if the divergence of $f_{\hat{\theta}}$ from f_{θ^*} is smaller than the divergence of f from f_{θ^*} for all local perturbations. Since equivocal observations are those whose likelihood ratios depend on the current beliefs, and thus the local nature of marginals-updating becomes salient, one way to interpret Condition **B2** is that it provides a measure of the frequency of equivocal observations and how heavily those observations get weighed *at a particular state* $\hat{\theta}$. We give a sufficient condition for **B2** in Lemma 11 in the appendix.

Given these conditions, beliefs need not converge asymptotically to the true state (i.e., learning may be inconsistent). A marginals-updating agent can be made to have arbitrarily high probability of asymptotically believing $\hat{\theta}$ is true by shifting the agent's initial beliefs very close to $\hat{\theta}$. We say that a state $\hat{\theta} \in \Theta$ may be *attracting* if, when prior beliefs are sufficiently close to $\hat{\theta}$, then posterior beliefs will converge to $\hat{\theta}$. While we focus on neighborhoods around $\hat{\theta}$ that in the limit converge to $\hat{\theta}$ almost surely, when the convergence probability is continuous then beliefs near $\hat{\theta}$ will with *positive* probability converge to $\hat{\theta}$. Lemma 7 provides conditions for continuity in the 2×2 model.

Theorem 2. *Let θ^* be the true state of the world and assume **B1**. Consider $\hat{\theta} \in \Theta$ such that for all i **B2** holds. Then for every $\varepsilon > 0$, there exists a $c > 0$ such that if $W_0\left(\theta_i = \hat{\theta}_i\right) \geq 1 - c$ for all i, then $\Pr\left(W_n\left(\theta_i = \hat{\theta}_i\right) \to 1\right) \geq 1 - \varepsilon$.*

In other words, a state $(\theta_1, \ldots, \theta_d)$ is attracting if perturbing one of the θ_i's into θ_i' while keeping the rest of the θ_{-i} static results in a very poor explanation of the world. In the 2×2 model, the state (L,H) could be attracting when the true state is (H,L) because (H,H) and (L,L), both one dimensional perturbations of (L,H), may not provide better explanations of the data. Furthermore, agents can use general copulas to combine marginal beliefs so long as the copula chosen is fixed or static (i.e., the correlation matrix does not update) and not sufficiently different from the independence copula (see the Appendix, assumption **B3**), giving the Theorem 2 a degree of robustness to the type of updating procedure used.

The basic premise of our inconsistency result Theorem 2 is that a state $\hat{\theta}$ only needs to be a better explanation for the real world than one-dimensional perturbations of $\hat{\theta}$ in order to be locally attracting. The following corollary then is a straightforward consequence of Gibb's inequality, stating that the true state of the world θ^* will have the local attraction property (an important sanity check):

Corollary 3. *Suppose also $f_{\theta^*} \notin \mathrm{co}(\{f_\theta : \theta \in G_i(\theta_0)\})$, as well as $|\Theta_i| < \infty$ for all i: then the belief $\theta = \theta^*$ is locally attracting in the sense of Theorem 2.*

Corollary 3 implies that an agent who mistakenly updates only along marginals may overcome this oversight by beginning with a prior which is close enough to the truth. In other words, if the beliefs about the state are close enough to correct, then the model with the misspecified copula is also close enough to correct so that the misspecification does not interfere with learning. This parallels a finding of Bohren and Hauser (2017) (Theorems 2 and 3), who find that, when agents have misspecified beliefs about the signal distribution, learning generally occurs so long as beliefs are "close enough" to the true probability.

4 Discussion and Implications

Our results show that whenever beliefs are multidimensional and subject to correlation neglect, and when information is not sufficiently rich to completely identify the model, belief heterogeneity is likely to persist or grow. We discuss the setup of our model and the interpretation of the results.

The reader may wonder about the simplicity of the model and whether our results are robust to more general setups. Our results are completely driven by the following assumptions: (i) a multidimensional model, with observations that do not completely identify the model, and (ii) rational, sequential updating of marginal beliefs but not the full joint distribution of beliefs. Our model is silent about where prior beliefs come from; perhaps there are behavioral or generational explanations for priors (see Bénabou and Tirole (2016) for an overview of belief production).

We have deliberately chosen the simplest model to illustrate that partial identification can lead to beliefs diverging from the truth—in fact, if anything the simplicity of our model and the set of signals make learning the truth more likely. There are many reasons to believe that the identification problem in a higher dimensional model with a richer signal-space would be more severe, making divergent learning even more likely, since equivocal observations would be even more common. Also, our agents can only possibly disagree about initial priors; as others have shown (e.g., Bohren and Hauser 2017) divergence would be even more likely if agents also disagreed about the model parameters.

A critical driver of our results is that agents update their marginal beliefs whenever they get new information. Learning is more likely to occur if information is more informative. Hence, "patient" agents who update their marginal priors only when $m \geq 2$ signals are observed in a row and together interpreted as one composite-signal are less susceptible to asymptotic inconsistency.

In reality people may desire to know the underlying state variables for reasons beyond (just) being able to predict future observations Y_n. For example, people may want to take actions whose payoffs depend on the values of θ and σ separately. Or there may be an additional variable whose realization will be made known in the far future, whose value could simply equal θ or σ. Thus, even if knowing $\theta + \sigma$ is sufficient to predict the distribution of Y_n (as is exactly true when the states are symmetric), knowledge of θ and σ separately would still matter. Nonetheless, marginal-updating agents need not converge to the same $\theta + \sigma$. Our simulation results presented in Appendix C show that when $\theta \neq \sigma$, beliefs may with positive probability converge to either (L, L) or (H, H) even though these are not the "symmetric" states.

Our model predicts that in some cases divergence may be unavoidable when initial beliefs are different. Initial heterogeneity can lead to complete polarization (in the sense of divergent beliefs), and more information worsens this outcome. One way to avoid divergence is for researchers to find ways to alleviate identification problems. However, if researchers produce only a handful of well-designed natural experiments or instrumental variables to relieve an identification problem, the people may nonetheless continue to hold divergent beliefs so long as there are plenty of cases in which a model leaves room for multiple interpretations.[5]

[5] We acknowledge Bruce Sacerdote for this observation.

5 Conclusion

The world is multidimensional—there are a number of factors that contribute to what we see—but the data we see are often lower dimensional than the world. Therefore we live with identification problems. It is also plausible that, rather than maintain a huge and costly joint distribution, people retain and update only marginal beliefs on the many variables which shape the signals people receive. However, Bayesian updating of marginal probabilities need not converge to the truth. We have characterized the limiting properties of beliefs for a simple model in which some observations do not identify the underlying parameters. Our main result is that when such observations are relatively more frequent, then asymptotically initial beliefs are likely to become reinforced. In particular, agents with differing priors may have posteriors diverge forever, with greater divergence the more common information received. However, if observations clearly identifying the model are relatively likely, beliefs converge to the truth with probability one, and divergence will not occur.

References

ACEMOGLU, D., V. CHERNOZHUKOV, AND M. YILDIZ (2016): "Fragility of asymptotic agreement under Bayesian learning," *Theoretical Economics*, 11, 187–225.

ANDRADE, P., R. K. CRUMP, S. EUSEPI, AND E. MOENCH (2016): "Fundamental disagreement," *Journal of Monetary Economics*, 83, 106–128.

ANDREONI, J. AND T. MYLOVANOV (2012): "Diverging opinions," *American Economic Journal: Microeconomics*, 209–232.

BALIGA, S., E. HANANY, AND P. KLIBANOFF (2013): "Polarization and ambiguity," *The American Economic Review*, 103, 3071–3083.

BAUMEISTER, C. AND J. D. HAMILTON (2015): "Sign Restrictions, Structural Vector Autoregressions, and Useful Prior Information," *Econometrica*, 83, 1963–1999.

BÉNABOU, R. AND J. TIROLE (2016): "Mindful economics: The production, consumption, and value of beliefs," *The Journal of Economic Perspectives*, 30, 141–164.

BENOÎT, J.-P. AND J. DUBRA (2018): "When do populations polarize? An explanation." Tech. rep., London Business School.

BERK, R. H. (1966): "Limiting behavior of posterior distributions when the model is incorrect," *The Annals of Mathematical Statistics*, 37, 51–58.

BOHREN, A. AND D. HAUSER (2017): "Bounded Rationality And Learning: A Framework and A Robustness Result," CEPR Discussion Papers 12036, C.E.P.R. Discussion Papers.

DIACONIS, P. AND D. FREEDMAN (1986): "On the Consistency of Bayes Estimates," *Ann. Statist.*, 14, 1–26.

DIXIT, A. K. AND J. W. WEIBULL (2007): "Political polarization," *Proceedings of the National Academy of Sciences*, 104, 7351–7356.

ENKE, B. AND F. ZIMMERMANN (2017): "Correlation neglect in belief formation," *forthcoming, Review of Economic Studies*.

ESPONDA, I. AND D. POUZO (2016): "Berk-Nash Equilibrium: A Framework for Modeling Agents With Misspecified Models," *Econometrica*, 84, 1093–1130.

EYSTER, E., A. GALEOTTI, N. KARTIK, AND M. RABIN (2014): "Congested observational learning," *Games and Economic Behavior*, 87, 519–538.

EYSTER, E. AND M. RABIN (2010): "Naive herding in rich-information settings," *American economic journal: microeconomics*, 2, 221–243.

FRYER, R. G., P. HARMS, AND M. O. JACKSON (2018): "Updating Beliefs when Evidence is Open to Interpretation: Implications for Bias and Polarization," *forthcoming, Journal of the European Economic Association*.

GABAIX, X. (2014): "A sparsity-based model of bounded rationality," *The Quarterly Journal of Economics*, 129, 1661–1710.

GERBER, A. AND D. GREEN (1999): "Misperceptions about perceptual bias," *Annual review of political science*, 2, 189–210.

HEIDHUES, P., B. KOSZEGI, AND P. STRACK (2015): "Unrealistic Expectations and Misguided Learning," *Available at SSRN*.

HIRSHLEIFER, D. AND S. H. TEOH (2003): "Limited attention, information disclosure, and financial reporting," *Journal of accounting and economics*, 36, 337–386.

HONG, H. AND J. C. STEIN (2007): "Disagreement and the stock market," *The Journal of Economic Perspectives*, 21, 109–128.

JERN, A., K.-M. K. CHANG, AND C. KEMP (2014): "Belief polarization is not always irrational." *Psychological review*, 121, 206.

KOMINERS, S. D., X. MU, AND A. PEYSAKHOVICH (2016): "Paying (for) Attention: The Impact of Information Processing Costs on Bayesian Inference," .

MALMENDIER, U., S. NAGEL, AND Z. YAN (2017): "The Making of Hawks and Doves: Inflation Experiences on the FOMC," Tech. rep., National Bureau of Economic Research.

ORTOLEVA, P. AND E. SNOWBERG (2015): "Overconfidence in political behavior," *The American Economic Review*, 105, 504–535.

RABIN, M. AND J. L. SCHRAG (1999): "First Impressions Matter: A Model of Confirmatory Bias," *The Quarterly Journal of Economics*, 114, 37–82.

SCHENK-HOPPÉ, K. R. AND B. SCHMALFUSS (2001): "Random fixed points in a stochastic Solow growth model," *Journal of Mathematical Economics*, 36, 19–30.

SCHWARTZSTEIN, J. (2014): "Selective Attention and Learning," *Journal of the European Economic Association*, 12, 1423–1452.

SETHI, R. AND M. YILDIZ (2016): "Communication with Unknown Perspectives," *Econometrica*, 84, 2029–2069.

SUNDARESAN, S. AND S. TURBAN (2014): "Inattentive valuation and belief polarization," Tech. rep., Working paper.

Appendix for Online Publication

A Proofs for the Simple Model

In the following lemma, for a sequence $\omega \in \{0,1,2\}^{\mathbb{N}}$, let ω_n denote the n^{th} element of ω:

Lemma 1. *If* $\lim_{n\to\infty} P_n$ *exists in* $[0,1]$ *almost everywhere and* $p_H, p_L \in (0,1)$*, then*

$$\Pr\left(\lim_{n\to\infty} P_n \in (0,1)\right) = 0. \tag{4}$$

Proof. Let $\Omega = \{0,1,2,\}^{\mathbb{N}}$ denote the set of all possible sequences of Y signals. Then $P_n \equiv P_n(\omega)$ can be written as a function of points $\omega \in \Omega$. Ω is endowed with the natural probability measure over signal observations and the σ-algebra generated by $P_n(\omega), Q_n(\omega)$. We claim that any point ω such that $\lim_{n\to\infty} P_n(\omega) \in (0,1)$ has $\omega_n \in \{0,2\}$ for only finitely many indices n. This is easy to see for $\omega_n = 2$: assume to the contrary that infinitely many $(n_m) \subset \mathbb{N}$ satisfy $\omega_{n_m} = 2$, and let $\lim_{n\to\infty} p_n(\omega) = c \in (0,1)$. Then taking such m arbitrarily high, $(P_{n_m+1}) \to \frac{cp_H}{cp_H+(1-c)p_L} > c$, a contradiction. The proof when $\omega_n = 0$ infinitely often is similar. (4) follows immediately by the Second Borel-Cantelli Lemma. □

Proof of Proposition 1. We prove for P_n (the proof for Q_n is similar). Suppose that $\theta = \sigma = H$. The discrete version of Corollary 2, using summation by parts instead of integration by parts in Theorem 1, establishes that P_n converges almost surely in $(0,1]$. Lemma 1 then implies that $P_n \overset{\text{a.s.}}{\to} 1$. The proof when $\theta = \sigma = L$ is similar. □

Proofs of Continuity Results

Consider the function $f(p,q)$ giving the probability of convergence of (P_n, Q_n) to (H,L) if $P_0 = p$, $Q_0 = q$ given p_H, p_L, q_H, q_L. This function has some nice properties. Suppose that $\theta = L, \sigma = H$. First,

$$\begin{aligned} f(p,q) =& p_L q_H f\left(P'((p,q),2), Q'((p,q),2)\right) + (1-p_L)(1-q_H)\left(P'((p,q),0), Q'((p,q),0)\right) \\ &+ \Big(p_L(1-q_H) + (1-p_L)q_H\Big) f\left(P'((p,q),1), Q'((p,q),1)\right). \end{aligned}$$

Furthermore, observe that for a *fixed* q, $f(p,q)$ is monotonically increasing in p:

Lemma 2. *For $p' \geq p$ and $q' \leq q$, $f(p',q') \geq f(p,q)$. Therefore, $f(\cdot,q)$ is continuous almost everywhere at any fixed q.*

Proof. Fix a history $\omega \in \Omega$. Let $P_n(\omega), Q_n(\omega)$ correspond to initial condition (p,q) and $P_n'(\omega), Q_n'(\omega)$ be defined to correspond to (p',q'). The claim can be verified by on the hypothesis that in each n, we have

$$P_n'(\omega) \geq P_n(\omega), \quad Q_n'(\omega) \leq Q_n(\omega). \tag{5}$$

The argument is accomplished with the observation that (i) If $Y_n(\omega) = 0, 2$, the ordering is preserved by monotonicity of the relevant functions. (ii) If $Y_n(\omega) = 1$, then a smaller value of Q_n corresponds with a larger increase in P_n. Conversely, a larger value of P_n corresponds with a smaller increase (larger decrease) in Q_n. This can be easily verified by noting that the odds ratio O_{n+1}^P increases by more when $Y_n = 1$ and Q_n is smaller. Conversely, O_{n+1}^Q increases by less when $Y_n = 1$ and P_n is larger. Because $O_{n+1}^P \geq O_{n+1}^{P'}$ if and only if $P_{n+1} \geq P_{n+1}'$, the inequalities in (5) are indeed preserved. □

In the Internet Appendix we provide a proof of more general features of continuity which builds on these results. In the proof, we discuss the relevance of the random dynamical theory of cocycles to our model. Lemma 7 is detailed in the internet appendix and shown to hold on a larger set of random dynamical systems with a Bernoulli shift as a random component. (Schenk-Hoppé and Schmalfuß, 2001) expands on our discussion of random dynamical systems via cocycles.

Proof of Proposition 2

There is a robust set of parameters such that beliefs diverge. Throughout suppose $\theta = L$ and $\sigma = H$. (By symmetry all results also hold for $\theta = H$ and $\sigma = L$.) First, we can characterize conditions on beliefs P, Q such that beliefs converge incorrectly almost surely. Ignoring a (measurably) small set of parameter values, we obtain a classification result for the asymptotic behavior of (P_n, Q_n). Lemma 3 states conditions on parameters that guarantee when priors exists such that beliefs will asymptotically converge to the endpoints (zero or one) with positive probability. Specifically, the proposition states that the asymptotic behavior of (P_n, Q_n) behavior can be stated in terms of the expectation of the transition function (i.e., the log odds ratio) in neighborhoods of the extremal points of $[p_L, p_H] \times [q_L, q_H]$, because asymptotically individuals will accumulate in these neighborhoods with positive probability.

Lemma 3. *The following hold:*

1. *If* $\mathbb{E}\left[\Delta\log O^P_{n+1}(q_L)\right] > 0$ *and* $\mathbb{E}\left[\Delta\log O^Q_{n+1}(p_H)\right] < 0$*, then there exist* (P_0, Q_0) *such that* $P_n \to 1$ *and* $Q_n \to 0$ *with positive probability tending to* 1 *as* $Q_0 \to 0$ *and* $P_0 \to 1$

2. *If* $\mathbb{E}\left[\Delta\log O^P_{n+1}(q_L)\right] < 0$*, then* $P \overset{a.s.}{\to} 0$*; if additionally* $\mathbb{E}\left[\Delta\log O^Q_{n+1}(p_L)\right] < 0$ *then* $Q \overset{a.s.}{\to} 0$*, whereas if* $\mathbb{E}\left[\Delta\log O^Q_{n+1}(p_L)\right] > 0$ *then* $Q \overset{a.s.}{\to} 1$*.*

3. *If* $\mathbb{E}\left[\Delta\log O^Q_{n+1}(p_H)\right] > 0$*, then* $Q \overset{a.s.}{\to} 1$*; if additionally* $\mathbb{E}\left[\Delta\log O^P_{n+1}(q_L)\right] < 0$ *then* $P \overset{a.s.}{\to} 0$*, whereas if* $\mathbb{E}\left[\Delta\log O^P_{n+1}(q_L)\right] > 0$ *then* $P \overset{a.s.}{\to} 1$*.*

Most importantly, if parameters satisfy condition 1 of Lemma 3 then there exist priors such that beliefs will diverge with positive probability. Numerically evaluating over all combinations $(p_H, p_L, q_H, q_L) \in (0,1)^4$, together with ordering restrictions, approximately 27.26% (Lebesgue measure) of parameters satisfy condition 1, meaning that for at least this many parameters there is a positive probability of converging to the wrong values for some priors.

Proof of Lemma 3

Consider the log-odds ratios, for which we have the following recursions:

$$\log O^P_{n+1}(2) = \log\frac{p_H}{p_L} + \log O^P_n \tag{6}$$

$$\log O^P_{n+1}(0) = \log\frac{1-p_H}{1-p_L} + \log O^P_n \tag{7}$$

$$\log O^P_{n+1}(1) = \log\frac{p_H(1-\overline{q}_n)+(1-p_H)\overline{q}_n}{p_L(1-\overline{q}_n)+(1-p_L)\overline{q}_n} + \log O^P_n. \tag{8}$$

Consider the random variable $\Delta\log O^P_{n+1}(\overline{q})$ (and the analogous expression for Q), whose expected value can be evaluated explicitly as:

$$\begin{aligned}\mathbb{E}\left[\Delta\log O^P_{n+1}(\overline{q})\right] =& p_L q_H \log\frac{p_H}{p_L} + (1-p_L)(1-q_H)\log\frac{1-p_H}{1-p_L} \\ &+ (p_L(1-q_H)+(1-p_L)q_H)\log\frac{p_H(1-\overline{q})+(1-p_H)\overline{q}}{p_L(1-\overline{q})+(1-p_L)\overline{q}}.\end{aligned}$$

The restriction of the transition function to the interior of the set $[p_L, p_H] \times [q_L, q_H]$ often is less relevant than its restriction to these endpoints.

We proceed in three lemmas, which deal with different cases in the parameters.

Lemma 4. *Let $\theta = L$ and $\sigma = H$. If $\mathbb{E}\left[\Delta \log O^P_{n+1}(q_L)\right] > 0$ and $\mathbb{E}\left[\Delta \log O^Q_{n+1}(p_H)\right] < 0$, then there exist (P_0, Q_0) such that $P_0 \to 1$ and $Q_0 \to 0$ with positive probability tending to 1 as $Q_0 \to 0$ and $P_0 \to 1$.*

Proof. The proof of this lemma will follow from Theorem 2, which is proved below. Specifically, the hypothesis of the lemma implies (3) holds with the true state of the world corresponding to $(\{\theta = L\}, \{\sigma = H\})$ and the false state t corresponding to $(\{\theta = H\}, \{\sigma = L\})$. □

Lemma 5. *If $\mathbb{E}\left[\Delta \log O^P_{n+1}(q_L)\right] < 0$, then $P \overset{a.s.}{\to} 0$, and if $\mathbb{E}\left[\Delta \log O^P_{n+1}(p_H)\right] > 0$, then $Q \overset{a.s.}{\to} 1$*

Proof. The proof follows the method of Lemma 4 by setting $Q^* = 1$ and $P^* = 0$. □

Lemma 6. *Let $\mathbb{E}\left[\Delta \log O^P_{n+1}(q_L)\right] < 0$. Then $P \overset{a.s.}{\to} 0$, and if $\mathbb{E}\left[\Delta \log O^Q_{n+1}(p_L)\right] < 0$ then $Q \overset{a.s.}{\to} 0$, whereas if $\mathbb{E}\left[\Delta \log O^Q_{n+1}(p_L)\right] > 0$ then $Q \overset{a.s.}{\to} 1$.*

Alternately, if $\mathbb{E}\left[\Delta \log O^Q_{n+1}(p_H)\right] > 0$, then $Q \overset{a.s.}{\to} 1$ and if $\mathbb{E}\left[\Delta \log O^P_{n+1}(q_L)\right] < 0$ then $P \overset{a.s.}{\to} 0$, whereas if $\mathbb{E}\left[\Delta \log O^P_{n+1}(q_L)\right] > 0$ then $P \overset{a.s.}{\to} 1$.

Proof. The convergence $P \overset{\text{a.s.}}{\to} 0$ follows from Lemma 5. As for convergence of Q, if $\mathbb{E}\left[\Delta \log O^P_{n+1}(p_L)\right] < 0$, then $\mathbb{E}\left[\Delta \log O^P_{n+1}(\overline{p})\right]$ is upper-bounded by a strictly negative number everywhere, and a law of large numbers argument suffices. On the other hand, if $\mathbb{E}\left[\Delta \log O^P_{n+1}(p_L)\right] > 0$, then as P concentrates almost surely at 0, $\mathbb{E}\left[\Delta \log O^Q_{n+1}\right]$ becomes lower bounded (for all sufficiently high n) by a positive $\varepsilon \in (0, \mathbb{E}\left[\Delta \log O^P_{n+1}(p_L)\right])$ almost surely, whence $Q \overset{\text{a.s.}}{\to} 1$. An obvious symmetry establishes the second claim. □

Next, given parameters p_H, p_L, q_H, q_L, let $f(p,q) = \Pr((P_n, Q_n) \to (1,0) | P_0 = p, Q_0 = q)$ be the probability of that beliefs converge to (H, L), which are the wrong beliefs. We have the following important results, including a finding on continuity in Lemma 7 that extends to a broader class of random dynamical systems with a Bernoulli Shift as a random component (see Proposition 1.6 in the Online Appendix).

Lemma 7. *If* $\frac{\log \frac{q_H}{q_L}}{\log \frac{p_H}{p_L}} \neq \frac{\log \frac{1-q_H}{1-q_L}}{\log \frac{1-p_H}{1-p_L}}$*, then the convergence probability* $f(p,q)$ *is continuous in priors* p,q*.*

The following Corollary is a related result which follows from Lemma 2.

Corollary 4. *If*

$$\frac{\log \frac{q_H}{q_L}}{\log \frac{p_H}{p_L}} < \frac{\log \frac{1-q_H}{1-q_L}}{\log \frac{1-p_H}{1-p_L}}, \tag{9}$$

then either $f(p,q) > 0$ *for all* $(p,q) \in (0,1) \times (0,1)$*, or it vanishes for all* (p,q)*.*

Thus, if parameters satisfy condition 1 of Lemma 3 as well as equation (9), then for any (rightly ordered) priors, beliefs will diverge with positive probability. The results of this section—namely, guaranteed divergence for symmetry together with continuity of the divergence probability—suggest that divergence is more likely to occur when the states (H,L) and (L,H) are more observationally similar (closer to symmetry). The more observationally similar are the states, the less the data can distinguish between competing beliefs when agents update only marginals. With exact observational equivalence, learning is guaranteed to converge to reinforce initial priors (this is also true with perfect Bayesian learning).

In our simulation results, we find that divergence is most likely to occur close to symmetry and decreases continuously. Compare these results to Fryer et al. (2018). In their model, there are two symmetric states (*a* and *b*) and agents' beliefs converge to one of these states to confirm their prior when ambiguous signals are sufficiently frequent. In their model, the probability of polarization is an increasing function of the probability of ambiguous signals. Our simulation results suggest that in our model the probability of divergence is a function of how different are parameters for each state. The more parameters differ, the easier it is to "statistically identify" observations of 1, which are otherwise unidentifiable. In their model ambiguous signals are completely unidentified. Thus, in our model, fundamentals determine the severity of the identification problem and the probability of divergence, whereas in their model the severity of ambiguity determines the probability of divergence. Furthermore, when states are not symmetric, posteriors can converge to values that are not even "symmetric" with the truth (i.e., to (H,H) or (L,L). This result is important because agents would not only disagree about the underlying values of θ, σ, but they would have quite different predictions for the distribution of Y_n. One could argue that polarization truly refers to agents disagreeing about the value of $\theta + \sigma$, which is what would occur in this case.

By evaluating the set of parameters which meet the bounds presented in Corollary 4 as well as those in Lemma 3, we find the our result. Approximately 8.33% (Lebesgue measure) of all parameters satisfy both sets of conditions. The sets of parameters for which (i) any priors will converge to (H,L) with positive probability, or for which (ii) some priors are guaranteed to converge to (H,L) with positive probability, indicate that divergence is most likely to occur in a neighborhood of parameters around symmetry.

Proof of Corollary 4. Note that

$$n \log \frac{p_H}{p_L} + \left\lceil -n \frac{\log \frac{q_H}{q_L}}{\log \frac{1-q_H}{1-q_L}} + 1 \right\rceil \log \frac{1-p_H}{1-p_L} = n \left(\log \frac{p_H}{p_L} + \left(-\frac{\log \frac{q_H}{q_L}}{\log \frac{1-q_H}{1-q_L}} + o(1) \right) \log \frac{1-p_H}{1-p_L} \right)$$

$$n \log \frac{q_H}{q_L} + \left\lceil -n \frac{\log \frac{q_H}{q_L}}{\log \frac{1-q_H}{1-q_L}} + 1 \right\rceil \log \frac{1-q_H}{1-q_L} < 0.$$

Hence, there must exist positive integers n, $m \approx -n \frac{\log \frac{q_H}{q_L}}{\log \frac{1-q_H}{1-q_L}}$ such that $n \log \frac{p_H}{p_L} + m \log \frac{1-p_H}{1-p_L} > 0$, and $n \log \frac{q_H}{q_L} + m \log \frac{1-q_H}{1-q_L} > 0$. It follows that if $Y = 2$ for $n\ell$ times and $Y = 0$ for $m\ell$ times, as ℓ becomes arbitrarily large, $\log O^P_{\ell(n+m)}$ becomes arbitrarily large and $\log O^Q_{\ell(n+m)}$ becomes arbitrarily small from any initial prior (P,Q). In particular, if f does not vanish for all (p,q), by Lemma 2, there is a critical O^{P*} and O^{Q*} such that for all pairs (O^P, O^Q) with $O^P \geq O^{P*}$ and $O^Q \leq O^{Q*}$, there is a positive probability that $O^P \to \infty$. Because this critical threshold can be reached in a finite number of steps from any prior, any prior has a positive probability of $O^P \to \infty$. The second claim is proved similarly by exchanging 'P' and 'Q'. □

Lemma 8. *Let $p_H = q_H$ and $p_L = q_L$. Then almost surely the ratio O^P/O^Q diverges to infinity if $P > Q$ and converges to zero if $P < Q$.*

Proof of Lemma 8. First, since $p_H = q_H$ and $p_L = q_L$, the likelihood ratios given observations

$Y = 0$ and $Y = 2$ are equal:

$$L_2^P = \frac{p_H}{p_L} = L_2^Q, \qquad L_0^P = \frac{1-p_H}{1-p_L} = L_0^Q.$$

Thus, O^P/O^Q is unchanged after these observations.

Second, O^P/O^Q increases after $Y = 1$ whenever $P > Q$. Differentiating $L_1^P(Q)$ with respect to Q shows that it is decreasing in Q. By symmetry, if $P > Q$, then $L_1^P(Q)/L_1^Q(P) > 1$. Similarly, differentiating and collecting terms, $L_1^P(Q)/L_1^Q(P)$ is increasing in P. Thus, as P/Q grows, the ratio of likelihoods grows and thus O^P/O^Q increases by more. Thus, $L_1^P(Q_n)/L_1^Q(P_n) \geq L_1^P(Q_0)/L_1^Q(P_0) > 1$, and so $L_1^P(Q_n)/L_1^Q(P_n)$ is bounded below by a number strictly greater than 1. Since all probabilities are strictly positive, by the Strong Law of Large Numbers asymptotically there will be an infinite number of $Y = 1$ observations, and thus O^P/O^Q increases without bound. By similar argument, if $P < Q$ then O^P/O^Q decreases to zero. When $p_H \neq q_H$ or $p_L \neq q_L$, the divergence of O^P/O^Q can occur with positive probability, but it is not guaranteed. □

Proof of Corollary 1. By Lemma 8 the ratio of the odds-ratios, O^P/O^Q converges to infinity almost surely. This implies that $P_n \overset{\text{a.s.}}{\to} 1$ almost surely. Since $O^P/O^Q \to \infty$, at least one of $P_n \overset{\text{a.s.}}{\to} 1$ or $Q_n \overset{\text{a.s.}}{\to} 0$. However, if $Q = 0$ then the problem of Bayesian learning is isomorphic to learning whether $\sigma = H$ given $P = 0$. That is, there exists $\varepsilon > 0$ such that if $P_n < \varepsilon$, then Q_n is a submartingale. By Doob's Martingale Convergence Theorem, this problem converges to the truth ($Q = 1$) almost surely (see Berk (1966) and Diaconis and Freedman (1986)). By symmetry, if $Q_n < \varepsilon$ then $P_n \overset{\text{a.s.}}{\to} 1$. □

The reader may wonder how much the 2 × 2 results depend on the specific form of the realization function $y(t,s)$. As stated, what matters for the possibility of inconsistency is the dimensionality of the realization function. Suppose instead that the function was: $Y = 0$ if $t = 0$ and $s = 1$; $Y = 2$ if $t = 1$ and $s = 0$; $Y = 1$ if t and s are equal. Then $Y = 1$ is still equivocal, but which states are difficult to learn have been "rotated." It is easy to see in light of these results that agents would learn the truth whenever $\theta \neq \sigma$, but agents would not necessarily learn the truth when $\theta = \sigma$. The difficulty now is that to distinguish between (H,H) and (L,L) agents need to have positively correlated beliefs, but beliefs are assumed to be independent.

B The General Model

Any W_n induces a probability distribution over the signal space $\mathbb{R}$ whose density w_n may be expressed as in the normal Bayesian case:

$$w_n\left(Y_{n+1}|\theta_i=\hat{\theta}_i\right)=\frac{\sum_{\vartheta\in\Theta:\vartheta_i=\hat{\theta}_i}W_n\left(\theta=\vartheta\right)f_{\vartheta}\left(Y_{n+1}\right)}{\sum_{\vartheta\in\Theta:\vartheta_i=\hat{\theta}_i}W_n\left(\theta=\vartheta\right)}, \tag{10}$$

$$w_n\left(Y_{n+1}|\theta_i\neq\hat{\theta}_i\right)=\frac{\sum_{\vartheta\in\Theta:\vartheta_i\neq\hat{\theta}_i}W_n\left(\theta=\vartheta\right)f_{\vartheta}\left(Y_{n+1}\right)}{\sum_{\vartheta\in\Theta:\vartheta_i\neq\hat{\theta}_i}W_n\left(\theta=\vartheta\right)}. \tag{11}$$

To update marginals, an agent reconstructs a joint distribution using the independence copula, denoted by Ψ, (i.e., an agent multiplies marginal probabilities) and then updates marginal distributions applying Bayes rule to marginal beliefs using the reconstructed joint distribution. Rigorously,

$$\begin{aligned}W_n\left(\theta=(\hat{\theta}_1,\ldots,\hat{\theta}_d)\right)=&\Psi\left(W_n^1,\ldots,W_n^d\right)(\hat{\theta}_1,\ldots,\hat{\theta}_d)\\ W_{n+1}\left(\theta_i=\hat{\theta}_i\right)\equiv&W_{n+1}\left(\theta_i=\hat{\theta}_i\middle|Y_{n+1}\right)\\ =&\frac{W_n\left(\theta_i=\hat{\theta}_i\right)w_n\left(Y_{n+1}\middle|\theta_i=\hat{\theta}_i\right)}{w_n\left(Y_{n+1}\right)}\\ =&\frac{W_n\left(\theta_i=\hat{\theta}_i\right)w_n\left(Y_{n+1}\middle|\theta_i=\hat{\theta}_i\right)}{W_n\left(\theta_i=\hat{\theta}_i\right)w_n\left(Y_{n+1}\middle|\theta_i=\hat{\theta}_i\right)+W_n\left(\theta_i\neq\hat{\theta}_i\right)w_n\left(Y_{n+1}\middle|\theta_i\neq\hat{\theta}_i\right)},\end{aligned}$$

(recall that for every $A\subset\Theta_i$ we have specified $W_n\left(\theta_i\in A\right)=\sum_{\hat{\theta}_i\in A}W_n\left(\theta_i=\hat{\theta}_i\right)$). The notation is used in lower-case with respect to Y to evoke that Y is continuously distributed with respect to a density, whereas θ is discretely valued with a probability mass function. For convenience, we will use $W_n^i(\cdot)=W_n(\theta_i=\cdot)$ interchangeably.

Note that the updating procedure implies

$$\log\frac{W_{n+1}\left(\theta_i=\hat{\theta}_i\right)}{1-W_{n+1}\left(\theta_i=\hat{\theta}_i\right)}=\log\frac{w_n\left(Y_{n+1}\middle|\theta_i=\hat{\theta}_i\right)}{w_n\left(Y_{n+1}\middle|\theta_i\neq\hat{\theta}_i\right)}+\log\frac{W_n\left(\theta_i=\hat{\theta}_i\right)}{1-W_n\left(\theta_i=\hat{\theta}_i\right)}. \tag{12}$$

By our probability shorthand $\sum_{\vartheta\in\Theta:\theta_i=\hat{\theta}_i}W_n(\theta=\vartheta)=W_n(\theta_i=\hat{\theta}_i)$ for every i. Thus, when $\sum_{\hat{\theta}_i\in\Theta_i}W_0(\theta_i=\hat{\theta}_i)=1$, one can verify that $\sum_{\hat{\theta}_i\in\Theta_i}W_n(\theta_i=\hat{\theta}_i)=1$ for every n, which is to say that we have a probability measure over marginals.

B.1 Approximate Optimality of the Independence Copula

The belief process of agents is boundedly rational. For Doob's well-known Bayesian consistency result to apply, agents must sequentially update their priors over the entire state space by calculating the joint posterior distribution after every observation. Our deviation from this benchmark can be rationalized by a sparsity-based model of bounded rationality following Gabaix (2014).

Consider d state variables that can take on k values. Then regular Bayesian updating requires storing k^d real numbers in the joint distribution. In practice agents might only maintain and update the marginal distributions over each dimension, thereby reducing the memory burden to kd real numbers. To update the marginal distributions from their observations and knowledge of the data generating process, the agents need to attempt to reconstruct the joint distribution of state variables. In this section we assume that agents apply the independence copula; our main result will allow agents substantially more latitude in the choice of a (fixed) copula.

Suppose that the agent is concerned with estimating a signal $Y_n \in \mathbb{R}^d$ as accurately as possible in the sense that he wishes to minimize a discounted square loss function

$$\min_{\{\hat{y}_n\}_{n=1}^{\infty}} \mathbb{E}\left[\sum_{n=1}^{\infty} \delta^{n-1} |Y_n - \hat{y}_n|^2\right].$$

The major behavioral assumption of this paper is that agents retain only their marginal beliefs $W_n^i, i = 1, \dots, d$ from period n. In the next period, in order to obtain W_{n+1}^i, the agents must reconstruct the joint distribution $W_n = W_n(\theta = (\hat{\theta}_1, \dots, \hat{\theta}_d))$ using some choice of copula which we have denoted Ψ. It is straightforward to see that the *independence copula*, given by

$$W_n(\theta = (\hat{\theta}_1, \dots, \hat{\theta}_d)) = \Psi\big(W_n^1, \dots, W_n^d\big)(\hat{\theta}_1, \dots, \hat{\theta}_d) = \prod_{i=1}^{d} W_n^i(\hat{\theta}_i), \tag{13}$$

correctly sets the initial joint prior W_0 equal to P, the distribution from which θ^* was sampled ex-ante. Hence, it is approximately optimal in the following sense:

Remark 1. *Suppose that $Y_n \in K \subset \mathbb{R}^d$ a.s., where K is a compact set. Let $\hat{y}_n^* = \mathbb{E}[y|W_n^*]$, where W_n^* is generated by application of the independence copula* (13) *in every period. Then for every*

$\varepsilon > 0$, *there is a* $\overline{\delta} \in (0,1)$ *such that for* $\delta \leq \overline{\delta}$,

$$\mathbb{E}\left[\sum_{n=1}^{\infty} \delta^{n-1}|Y_n - \hat{y}_n^*|^2\right] \leq \inf_{\{\hat{y}_n\}_{n=1}^{\infty} \subset K} \mathbb{E}\left[\sum_{n=1}^{\infty} \delta^{n-1}|Y_n - \hat{y}_n|^2\right] + \varepsilon.$$

Thus, for a sufficiently impatient agent, the strategy of retaining only marginals and reconstructing joint distributions via multiplication may be arbitrarily close to optimal.

Proof of Remark 1. First, note that in period $n = 1$, $\hat{y}_n^* = \mathbb{E}[y]$, where the expectation is over the prior probability distribution $Q_1 \times \cdots \times Q_d$ on Θ and on the contingent distribution of Y induced by θ. Hence, $\mathbb{E}\left[|Y_n - \hat{y}_n^*|^2\right] = \mathbb{E}\left[|y - \mathbb{E}[y]|^2\right] \leq \mathbb{E}\left[|Y_n - \hat{y}_n|^2\right]$ for all r.v. $\hat{y}_n$. The desired result follows from noting that $\mathbb{E}\left[\sum_{n=1}^{\infty} \delta^{n-1}|Y_n - \hat{y}_n^*|^2\right] \leq \mathbb{E}\left[|Y_n - \hat{y}_n^*|^2\right] + \frac{\delta C}{1-\delta}$ for some constant $C < \infty$ depending on the set K. □

For instance, the independence copula is also approximately optimal in the following sense, which may be verified by induction:

Lemma 9. *Suppose that the probability density function $f_\theta(y)$ is multiplicatively separable in the sense that for every $\theta \in \Theta$, $f_\theta(y) = \prod_{i=1}^{d} f_{\theta_i}^i(y)$ for some functions $f_{\theta_i}^i$, $i \in \{1, \ldots, d\}$, $\theta_i \in \Theta_i$. Then when W_n^* is generated by application of the independence copula it exactly equals the canonical Bayesian prior conditioning on $Y_1, \ldots, Y_{n-1}$ and initial prior $W_0 =$* P.

This covers the obvious situation where Y is d-dimensional with each coordinate independently chosen from distribution $f_{\theta_i}^i$. So when the signal Y is extremely well-identified, updating along marginals is equivalent to canonical Bayesian updating and is consistent under light conditions.

B.2 Proofs

As will be evident in the proof, assumption **A3** is only necessary insofar as it provides perturbations to marginal beliefs in the interior of the d dimensional unit cube: in this respect, it may be generalized a number of ways. Lemma 1 is an example of such a perturbation argument.

Proof of Theorem 1. Recall that $Y_n \overset{\text{iid}}{\sim} F_{\theta^*}$. When the independence copula is applied, the denominator of (10) is

$$\sum_{\vartheta\in\Theta:\vartheta_i=\theta_i^*} W_n(\theta=\vartheta) = \sum_{\vartheta\in\Theta:\vartheta_i=\theta_i^*} W_n(\theta_i=\theta_i^*)\prod_{j\neq i} W_n(\theta_j=\vartheta_j)$$
$$= W_n(\theta_i=\theta_i^*)\sum_{\vartheta_{-i}\in\Theta_{-i}}\prod_{j\neq i} W_n(\theta_j=\vartheta_j) = W_n(\theta_i^*).$$

Expanding the denominator of (11) in a similar fashion allows us to derive the following conditional expectation:

$$\mathbb{E}_n\left(W_{n+1}(\theta_i^*)\right) = W_n(\theta_i=\theta_i^*)\int_{\mathbb{R}} \frac{w_n(y|\theta_i=\theta_i^*)}{w_n(y)} f_{\theta^*}(y)\,\mathrm{d}y$$
$$= W_n(\theta_i=\theta_i^*)\int_{\mathbb{R}} \frac{\sum_{\theta:\theta_i=\theta_i^*} f_\theta(y)\prod_{j\neq i} W_n(\theta_j)}{\sum_{\theta\in\Theta} f_\theta(y)\prod_{j=1}^d W_n(\theta_j)} f_{\theta^*}(y)\,\mathrm{d}y.$$

Integration by parts implies

$$\int_{\mathbb{R}} \frac{\sum_{\theta:\theta_i=\theta_i^*} f_\theta(y)\prod_{j\neq i} W_n(\theta_j)}{\sum_{\theta\in\Theta} f_\theta(y)\prod_{j=1}^d W_n(\theta_j)} f_{\theta^*}(y)\,\mathrm{d}y - 1$$
$$= \int_{\mathbb{R}} \frac{f_{\theta^*}(y)}{\sum_{\theta\in\Theta} f_\theta(y)\prod_{j=1}^d W_n(\theta_j)}\left(\sum_{\theta:\theta_i=\theta_i^*} f_\theta(y)\prod_{j\neq i} W_n(\theta_j) - \sum_{\theta\in\Theta} f_\theta(y)\prod_{j=1}^d W_n(\theta_j)\right)\mathrm{d}y$$
$$= \int_{\mathbb{R}}\left[\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{f_{\theta^*}(y)}{\sum_{\theta\in\Theta} f_\theta(y)\prod_{j=1}^d W_n(\theta_j)}\right)\right.$$
$$\left.\int_y^\infty \left(\sum_{\theta:\theta_i=\theta_i^*} f_\theta(z)\prod_{j\neq i} W_n(\theta_j) - \sum_{\theta\in\Theta} f_\theta(z)\prod_{j=1}^d W_n(\theta_j)\right)\mathrm{d}z\right]\mathrm{d}y,$$

where in the third line we have used that the term in parenthesis on the preceding line is a difference of two probability densities on $\mathbb{R}$. By A1

$$\frac{f'_{\theta^*}(y)}{f_{\theta^*}(y)} \geq \max_{\theta\in\Theta}\frac{f'_\theta(y)}{f_\theta(y)} \geq \frac{\sum_{\theta\in\Theta} f'_\theta(y)\prod_{j=1}^d W_n(\theta_j)}{\sum_{\theta\in\Theta} f_\theta(y)\prod_{j=1}^d W_n(\theta_j)},$$

so the quotient rule implies $\frac{\mathrm{d}}{\mathrm{dy}}\left(\frac{f_{\theta^*}(y)}{\sum_{\theta\in\Theta} f_\theta(y)\prod_{j=1}^d W_n(\theta_j)}\right) \geq 0$. On the other hand, one can write

$$\begin{aligned}
&\sum_{\theta:\theta_i=\theta_i^*} f_\theta(z)\prod_{j\neq i} W_n(\theta_j) - \sum_{\theta\in\Theta} f_\theta(z)\prod_{j=1}^d W_n(\theta_j)\\
&\quad= \sum_{\theta:\theta_i=\theta_i^*} f_\theta(z)\prod_{j\neq i}W_n(\theta_j) - W_n(\theta_i^*)\sum_{\theta:\theta_i=\theta_i^*} f_\theta(z)\prod_{j\neq i}W_n(\theta_i) - \sum_{\theta:\theta_i\neq\theta_i^*} f_\theta(z)\prod_{j=1}^d W_n(\theta_j)\\
&\quad=(1-W_n(\theta_i^*))\sum_{\theta:\theta_i=\theta_i^*} f_\theta(z)\prod_{j\neq i}W_n(\theta_j) - \sum_{\theta_{-i}\in\Theta_{-i}}\prod_{j\neq i}W_n(\theta_j)\sum_{\theta_i\in\Theta_i\setminus\{\theta_i^*\}} f_\theta(z)W_n(\theta_i)\\
&\quad=(1-W_n(\theta_i^*))\sum_{\substack{\theta_{-i}\in\Theta_{-i}\\ \theta_i=\theta_i^*}} f_\theta(z)\prod_{j\neq i}W_n(\theta_j)\\
&\qquad-(1-W_n(\theta_i^*))\sum_{\theta_{-i}\in\Theta_{-i}}\prod_{j\neq i}W_n(\theta_j)\sum_{\theta_i\in\Theta_i\setminus\{\theta_i^*\}} f_\theta(z)\frac{W_n(\theta_i)}{(1-W_n(\theta_i^*))}.
\end{aligned}$$

Fix a $\theta_{-i}\in\Theta_{-i}$ and let θ^s denote the element $\theta\in\Theta$ with $\theta^s_{-i}=\theta_{-i}$ and $\theta_i^s=\theta_i^*$. Let $S\subset\Theta$ denote the set of ϑ with $\vartheta_{-i}=\theta_{-i}$ and $\vartheta_i\neq\theta_i^*$. **A2** implies that for any $|S|$-dimensional probability vector $\mathbf{q}$ over S and any $y\in\mathbb{R}$,

$$\int_y^\infty \left(f_{\theta^s}(z) - \left\langle (f_\vartheta(z))_{\vartheta\in S}, \mathbf{q}\right\rangle\right)\mathrm{d}z \geq 0.$$

Because $\left(\frac{W_n(\theta_i)}{1-W_n(\theta_i^*)}\right)_{\theta_i\neq\theta_i^*}$ is one such probability vector, it follows that

$$\int_y^\infty \sum_{\theta:\theta_i=\theta_i^*} f_\theta(z)\prod_{j\neq i}W_n(\theta_j) - \sum_{\theta\in\Theta} f_\theta(z)\prod_{j=1}^d W_n(\theta_j)\,\mathrm{d}z \geq 0.$$

Hence, $W_n(\theta_i^*)$ is indeed a submartingale for all i. As it is bounded between 0 and 1, Doob's convergence theorem for submartingales establishes the first claim (Berk, 1966). It can similarly be shown that the inverse odds ratio $\frac{1-W_n(\theta_i^*)}{W_n(\theta_i^*)}$ is a nonnegative supermartingale. In particular, $\mathbb{E}_0\left(\frac{1-W_n(\theta_i^*)}{W_n(\theta_i^*)}\right)$ is at least weakly decreasing, so almost surely $\lim_{n\to\infty} W_n(\theta_i^*) > 0$.

Now, suppose that A1–A3 hold. Suppose for the sake of contradiction that for $i=1,\ldots,d$,

$\lim_{n\to\infty} W_n(\theta_i^*) = L_i$, where for some i it is the case that $L_i < 1$. Fixing this i, note that for any n

$$\frac{w_n(y|\theta_i = \theta_i^*)}{w_n(y|\theta_i \neq \theta_i^*)}$$

$$= \frac{\sum_{\substack{\theta_{-i}\in\Theta_{-i}\setminus\{\theta^*_{-i}\} \\ \theta_i=\theta_i^*}} f_\theta(y)\prod_{j\neq i} W_n(\theta_j) + f_{\theta^*}(y)\prod_{j\neq i} W_n(\theta_j^*)}{\sum_{\theta_{-i}\in\Theta_{-i}\setminus\{\theta^*_{-i}\}} \prod_{j\neq i} W_n(\theta_j) \sum_{\theta_i\in\Theta_i\setminus\{\theta_i^*\}} f_\theta(y)\frac{W_n(\theta_i)}{(1-W_n(\theta_i^*))} + \sum_{\theta_i\in\Theta_i\setminus\{\theta_i^*\}} f_{(\theta_i,\theta^*_{-i})}(y)\frac{W_n(\theta_i)\prod_{j\neq i} W_n(\theta_j^*)}{(1-W_n(\theta_i^*))}}.$$

By restricting attention to $A\bigcap_{\substack{\theta:\theta_i=\theta_i^* \\ \theta_{-i}\neq\theta^*_{-i}}}\{y : f_\theta(y) < M\}$ if necessary, for large enough M, we may assume that $f_\theta(y)$ for θ satisfying $\theta_i = \theta_i^*$ and $\theta_{-i} \neq \theta^*_{-i}$ is bounded by some M on A. Similarly, we may assume that $f_{\theta^*} > f_\theta + \varepsilon$ for some $\varepsilon > 0$ for $y \in A$. Using **A3** and the arguments used to prove convergence, $y \in A$ implies

$$\begin{aligned}\sum_{\theta_{-i}\in\Theta_{-i}\setminus\{\theta^*_{-i}\}} \prod_{j\neq i} W_n(\theta_j) \sum_{\theta_i\in\Theta_i\setminus\{\theta_i^*\}} f_\theta(y)\frac{W_n(\theta_i)}{(1-W_n(\theta_i^*))} &\leq \sum_{\substack{\theta_{-i}\in\Theta_{-i}\setminus\{\theta^*_{-i}\} \\ \theta_i=\theta_i^*}} f_\theta(y)\prod_{j\neq i} W_n(\theta_j) \\ &\leq M\left(1-\prod_{j\neq i} W_n(\theta_j^*)\right) < \infty.\end{aligned}$$

In addition,

$$\begin{aligned}\lim_{n\to\infty} f_{\theta^*}(y)\prod_{j\neq i} W_n(\theta_j^*) = f_{\theta^*}(y)\prod_{j\neq i} L_j &\geq \sum_{\theta_i\in\Theta_i\setminus\{\theta_i^*\}} f_{(\theta_i,\theta^*_{-i})}(y)\frac{W_n(\theta_i)\prod_{j\neq i} L_j}{(1-W_n(\theta_i^*))} + \varepsilon\prod_{j\neq i} L_j \\ &= \lim_{n\to\infty} \sum_{\theta_i\in\Theta_i\setminus\{\theta_i^*\}} f_{(\theta_i,\theta^*_{-i})}(y)\frac{W_n(\theta_i)\prod_{j\neq i} W_n(\theta_j^*)}{(1-W_n(\theta_i^*))} + \varepsilon\prod_{j\neq i} W_n(\theta_j^*)\end{aligned}$$

By the second Borel-Cantelli Lemma, $\Pr(\limsup_{n\to\infty}\{Y_n \in A\}) = 1$. Notice that the first term in the previous line is bounded by $M\prod_{j\neq i} W_n(\theta_j^*)$. Hence, almost surely

$$\limsup_{n\to\infty} \frac{w_n(y|\theta_i = \theta_i^*)}{w_n(y|\theta_i \neq \theta_i^*)} \geq \frac{M+\varepsilon\prod_{j\neq i} L_j}{M} > 1,$$

since $L_j > 0$ for all j. Recall that

$$\frac{1-W_{n+1}(\theta_i^*)}{W_{n+1}(\theta_i^*)} = \frac{w_n(Y_{n+1}|\theta_i = \theta_i^*)}{w_n(Y_{n+1}|\theta_i \neq \theta_i^*)}\,\frac{1-W_n(\theta_i^*)}{W_n(\theta_i^*)},$$

so it is impossible that the inverse odds ratio converges. Hence, $L_i = 1$ for all i almost surely. □

Proof of Corollary 2. It suffices to verify that assumptions **A1** and **A2** hold. We do this by verifying that for every θ and every i, $\frac{f_{T_i(\theta)}(y)}{f_\theta(y)}$ is weakly increasing. Suppose without loss of generality that $i = d$. By the convolution formula and our hypothesis,

$$\begin{aligned}
f'_{T_d(\theta)}(y)f_\theta(y) =& \frac{\partial}{\partial y}\int\cdots\int g_{\theta_1}(x_1)\cdots g_{\theta_d^*}(y-x_1-\cdots x_{d-1})\,\mathrm{d}x_1\cdots\mathrm{d}x_{d-1}\\
&\cdot\int\cdots\int g_{\theta_1}(x'_1)\cdots g_{\theta_d}(y-x'_1-\cdots x'_{d-1})\,\mathrm{d}x'_1\cdots\mathrm{d}x'_{d-1}\\
=&\int\cdots\int g_{\theta_1}(x_1)g_{\theta_1}(x'_1)\cdots g'_{\theta_d^*}(y-x_1-\cdots x_{d-1})g_{\theta_d}(y-x'_1-\cdots x'_{d-1})\,\mathrm{d}x_1\cdots\mathrm{d}x'_{d-1}\\
\geq&\int\cdots\int g_{\theta_1}(x_1)g_{\theta_1}(x'_1)\cdots g_{\theta_d^*}(y-x_1-\cdots x_{d-1})g'_{\theta_d}(y-x'_1-\cdots x'_{d-1})\,\mathrm{d}x_1\cdots\mathrm{d}x'_{d-1}\\
=&f_{T_d(\theta)}(y)f'_\theta(y).
\end{aligned}$$

□

Lemma 10. *Let $\{f_n\}_{n=0}^\infty$ be a collection of probability densities on Y satisfying*

$$0 < \beta_1 \equiv \inf_{n,y}\frac{f_n(y)}{f_0(y)} \leq \sup_{n,y}\frac{f_n(y)}{f_0(y)} \equiv \beta_2 < \infty.$$

Then for any vector of nonnegative weights $(p_n)_{n=0}^\infty$ and any $n_0 \in \mathbb{N}$ satisfying $\sum_{k=0}^\infty p_k = 1$,

$$\begin{aligned}
&\left|D_{KL}\left(f_0\,\middle\|\,\sum_{k=0}^{n_0-1}p_kf_k + f_1\sum_{k=n_0}^\infty p_k\right) - D_{KL}\left(f_0\,\middle\|\,\sum_{k=0}^\infty p_kf_k\right)\right| && (14)\\
&\quad\leq \max\left\{\log\frac{\beta_1}{\beta_1-(\beta_2-\beta_1)\sum_{k=n_0}^\infty p_k}, -\log\frac{\beta_1}{\beta_1+(\beta_2-\beta_1)\sum_{k=n_0}^\infty p_k}\right\}. && (15)
\end{aligned}$$

Proof. This follows from the calculation:

$$\begin{aligned}
&\left|\int\left(\log\frac{f_0}{\sum_{k=0}^{n_0-1}p_kf_k+f_1\sum_{k=n_0}^{\infty}p_k}-\log\frac{f_0}{\sum_{k=0}^{\infty}p_kf_k}\right)f_0\right|\mathrm{d}y\\
&\quad\leq\int\left|\log\frac{\sum_{k=0}^{\infty}p_kf_k}{\sum_{k=0}^{n_0-1}p_kf_k+f_1\sum_{k=n_0}^{\infty}p_k}\right|f_0\,\mathrm{d}y. \qquad (16)
\end{aligned}$$

where

$$\begin{aligned}
\log\frac{\sum_{k=0}^{\infty}p_kf_k}{\sum_{k=0}^{n_0-1}p_kf_k+f_1\sum_{k=n_0}^{\infty}p_k}&\leq\log\frac{\sum_{k=0}^{\infty}p_kf_k}{\sum_{k=0}^{\infty}p_kf_k-(\beta_2-\beta_1)\sum_{k=n_0}^{\infty}p_kf_0}\\
&\leq\log\frac{\beta_1}{\beta_1-(\beta_2-\beta_1)\sum_{k=n_0}^{\infty}p_k} \qquad (17)
\end{aligned}$$

and similarly

$$\log\frac{\sum_{k=0}^{\infty}p_kf_k}{\sum_{k=0}^{n_0-1}p_kf_k+f_1\sum_{k=n_0}^{\infty}p_k}\geq\log\frac{\beta_1}{\beta_1+(\beta_2-\beta_1)\sum_{k=n_0}^{\infty}p_k} \qquad (18)$$

Since $\int f_0\,\mathrm{d}y=1$, (16) is bounded in magnitude by the larger of the magnitudes of (17) and (18).

□

Inconsistent learning

In the main text we assume that agents us an independence copula to construct joint probabilities given marginals. However, we can use more general (fixed, static) copulas so long as the copula is well behaved. We make the following assumptions on the copula Ψ which are that the agent updates using roughly the independence cupola, and that no realization of Y rule out any state θ:

B3: The copula Ψ is relatively bounded with respect to the independence copula;

$$0<\inf_{\substack{t\in\Theta\\V^i\in\Delta(\Theta_i),\\1\leq i\leq d}}\frac{\Psi\big(V^1,\ldots,V^d\big)(t_1,\ldots,t_d)}{\prod_{i=1}^{d}V^i(\hat{\theta}_i)}\leq\sup_{\substack{t\in\Theta\\V^i\in\Delta(\Theta_i),\\1\leq i\leq d}}\frac{\Psi\big(V^1,\ldots,V^d\big)(t_1,\ldots,t_d)}{\prod_{i=1}^{d}V^i(\hat{\theta}_i)}<\infty. \qquad (19)$$

It is clear that the independence copula of the main text satisfies **B3**.

One can illustrate the substance of Assumption **B2** by rephrasing it in terms of the common L^1 distance between probability densities, $\|f-g\|_1 \equiv \int_X |f-g|\,\mathrm{d}\mu$. In this case the local dominance condition (3) states that $f_{\hat{\theta}}$ is "closer" to f_{θ^*} than any of the $f \in \mathrm{co}\left(f_\theta : \theta \in G_i(\hat{\theta})\right)$ are to $f_{\hat{\theta}}$, which is essentially to state that $f_{\hat{\theta}}$ is closer to f_{θ^*} than any of the perturbed f. Connecting the K-L divergence to an L^1 norm requires assumption B1, although the bound in (2) only enters inside of a logarithm.

Lemma 11. *Suppose that **B1** holds with bound β. If $|\Theta_i| < \infty$ then **B2** holds if and only if $D_{KL}\left(f_{\theta^*}\,||f\,\right) > D_{KL}\left(f_{\theta^*}\,\left|\right|f_{\hat{\theta}}\,\right)$ for all $f \in \mathrm{co}\left(\left\{f_\theta : \theta \in G_i(\hat{\theta})\right\}\right)$. Moreover a sufficient condition for **B2** is*

$$\left\|f_{\hat{\theta}} - f_{\theta^*}\right\|_1 < \frac{1}{2\log\beta} \inf_{f\in\mathrm{co}\left(\left\{f_\theta\in G_i(\hat{\theta})\right\}\right)} \|f - f_{\theta^*}\|_1^2.$$

Proof of Lemma 11. Necessity of the first condition is clear. To prove sufficiency, suppose for the sake of contradiction that $\inf_{f\in\mathrm{co}\left(\left\{f_\theta:\theta\in G_i(\hat{\theta})\right\}\right)} D_{\mathrm{KL}}\left(f_{\theta^*}\,||f\,\right) = D_{\mathrm{KL}}\left(f_{\theta^*}\,\left|\right|f_{\hat{\theta}}\,\right)$; then there is a sequence of weights $p^m \equiv (p_1,\ldots,p_{|\Theta_i|-1})_m \in \Delta(G_i(t))$ for all i such that, letting f^m denote the element of $\mathrm{co}\left(\left\{f_\theta : \theta \in G_i(\hat{\theta})\right\}\right)$ with weights p^m, satisfies $D_{\mathrm{KL}}\left(f_{\theta^*}\,||f_m\,\right) \to D_{\mathrm{KL}}\left(f_{\theta^*}\,\left|\right|f_{\hat{\theta}}\,\right)$. By a diagonalization argument we can extract a limiting probability vector p and a subsequence p^{m_ℓ} such that $p^{m_\ell} \to p$ pointwise. But if f is the element of $\mathrm{co}\left(\left\{f_\theta : \theta \in G_i(\hat{\theta})\right\}\right)$ corresponding to the weight vector p, the dominated convergence theorem (invoking B3) implies that $D_{\mathrm{KL}}\left(f_{\theta^*}\,||f\,\right) = D_{\mathrm{KL}}\left(f_{\theta^*}\,\left|\right|f_{\hat{\theta}}\,\right)$, a contradiction to our first stated hypothesis. Note that Prokhorov's theorem generalizes the same argument to sets of densities $\{f_\theta\}_{\theta\in\Theta}$ that are tight with closed convex hull.

To prove the second claim, first we claim that for all densities f, g we have $D_{\mathrm{KL}}\left(f\,||g\,\right) \geq \frac{1}{2}\|f-g\|_1^2$. This is just a continuous version of Pinsker's inequality, which we give here for the L^1

norm. Let $a = \int_{\{f>g\}} f \, \mathrm{d}\mu$ and $b = \int_{\{f>g\}} g \, \mathrm{d}\mu$ (ignore the case that $\|f-g\|_1 = 0$):

$$\begin{aligned} -D_{\mathrm{KL}}\left(f \,\|g\,\right) = \int_X f \log \frac{g}{f} \, \mathrm{d}\mu &= a \int_{\{f>g\}} \frac{f}{a} \log \frac{g}{f} \, \mathrm{d}\mu + (1-a) \int_{\{g \geq f\}} \frac{f}{1-a} \log \frac{g}{f} \, \mathrm{d}\mu \\ &\leq a \log \frac{b}{a} + (1-a) \log \frac{1-b}{1-a} \\ &\leq -2(a-b)^2, \end{aligned}$$

where the second line uses Jensen's inequality and the last line is an application of the fact that $a \log \frac{b}{a} + (1-a) \log \frac{1-b}{1-a} + 2(a-b)^2$ achieves a maximum at $a = b$. Note that

$$\int |f-g| \, \mathrm{d}\mu = \int_{\{f>g\}} (f-g) \, \mathrm{d}\mu - \int_{\{f \leq g\}} (f-g) \, \mathrm{d}\mu = 2 \int_{f>g} (f-g) \, \mathrm{d}\mu,$$

and so $D_{\mathrm{KL}}\left(f \,\|g\,\right) \geq 2 \left(\int_{\{f>g\}} (f-g) \, \mathrm{d}\mu \right)^2 = \frac{1}{2} \|f-g\|_1^2$. By B3, for any $f \in \mathrm{co}\left(\{f_\theta \in G_i(t)\}\right)$

$$\begin{aligned} D_{\mathrm{KL}}\left(f_{\theta^*} \,\|f_{\hat{\theta}}\,\right) \leq D_{\mathrm{KL}}\left(f_{\theta^*} \,\|f_{\hat{\theta}}\,\right) + D_{\mathrm{KL}}\left(f_{\hat{\theta}} \,\|f_{\theta^*}\,\right) &= \int (f_{\theta^*} - f_{\hat{\theta}}) \log \frac{f_{\theta^*}}{f_{\hat{\theta}}} \, \mathrm{d}\mu \\ &\leq \sup_x \log \frac{f_{\theta^*}}{f_{\hat{\theta}}} \int |f_{\theta^*} - f_{\hat{\theta}}| \, \mathrm{d}\mu \leq \log \beta \, \|f_t - f_{\theta^*}\|_1 . \end{aligned}$$

By Pinsker's inequality we have thus shown

$$\begin{aligned} D_{\mathrm{KL}}\left(f_{\theta^*} \,\|f_{\hat{\theta}}\,\right) \leq \log \beta \, \|f_{\theta^*} - f_{\hat{\theta}}\|_1 &< \frac{1}{2} \inf_{f \in \mathrm{co}(\{f_\theta \in G_i(t)\})} \|f_{\theta^*} - f\|_1^2 \\ &\leq \inf_{f \in \mathrm{co}(\{f_\theta \in G_i(t)\})} D_{\mathrm{KL}}\left(f_{\theta^*} \,\|f\,\right). \end{aligned}$$

□

Proof of Theorem 2. Denote lower and and upper limits in (19) by α_1 and α_2 respectively, and let the left side of (2) be β. For any fixed i, we claim that there is a $\delta \in (0,1)$ such that if

$$W_n^j(\hat{\theta}_j) \geq \delta \text{ for all } j \tag{20}$$

then $\mathbb{E}\left[\log\frac{w_n\left(Y_{n+1}\middle|\theta_i=\hat{\theta}_i\right)}{w_n\left(Y_{n+1}\middle|\theta_i\neq\hat{\theta}_i\right)}\right]>0$. First, denote

$$L\equiv\inf_{f\in\mathrm{co}(\{f_\theta:\theta\in G_i(t)\})}D_{\mathrm{KL}}\left(f_{\theta^*}\,||f\,\right)>D_{\mathrm{KL}}\left(f_{\theta^*}\,\middle|\middle|f_{\hat{\theta}}\,\right).\tag{21}$$

Now by (10), $w_n\left(Y_{n+1}|\theta_i=\hat{\theta}_i\right)$ is a weighted average of $f_\vartheta\left(Y_{n+1}\right)$ such that $\vartheta_i=\hat{\theta}_i$; furthermore, by (19) one has $W_n(\theta=\hat{\theta})\geq\alpha_1\prod_{j=1}^d W_n(\hat{\theta}_j)$ and:

$$\sum_{\substack{\vartheta\in\Theta:\vartheta_i=\hat{\theta}_i\\ \vartheta_{-i}\neq t_{-i}}}W_n(\theta=\vartheta)\leq\sum_{j\neq i}\sum_{\substack{\vartheta\in\Theta:\vartheta_i=\hat{\theta}_i\\ \vartheta_j\neq\hat{\theta}_j}}W_n(\theta=\vartheta)\leq\sum_{j\neq i}(1-W_n^j(\hat{\theta}_j)).$$

From (11), $w_n(Y_{n+1}|\theta_i\neq\hat{\theta}_i)$ can similarly be expressed as a weighted average with weights satisfying

$$\sum_{\vartheta\in\Theta:\vartheta_i\neq\hat{\theta}_i}W_n(\theta=\vartheta)=\sum_{s\in\Theta_i:s\neq\hat{\theta}_i}\left(W_n(\theta_i=s,\theta_{-i}=\hat{\theta}_{-i})+\sum_{\substack{\vartheta\in\Theta:\theta_i=s\\ \theta_{-i}\neq\hat{\theta}_{-i}}}W_n(\theta=\vartheta)\right).$$

By (19), $W_n(\theta_i=s,\theta_{-i}=\hat{\theta}_{-i})\geq\alpha_1W_n^i(s)\prod_{j\neq i}W^j(\hat{\theta}_j)$ and

$$\begin{aligned}\sum_{\substack{\vartheta\in\Theta:\theta_i=s\\ \theta_{-i}\neq t_{-i}}}W_n(\theta=\vartheta)&\leq\alpha_2W_n^i(s)\sum_{j\neq i}\sum_{\vartheta\in\Theta_{-i}:\vartheta_j\neq\hat{\theta}_j}\prod_{k\neq i}W_n^k(\vartheta_k)\\ &=\alpha_2W_n^i(s)\sum_{j\neq i}(1-W_n^j(\hat{\theta}_j)).\end{aligned}$$

Thus, in calculating (10) the agent places little weight on ϑ that are not $\hat{\theta}$, and in calculating (11), the agent places little weight on ϑ such that $\vartheta_{-i}\neq\hat{\theta}_{-i}$. In particular, the weight placed on densities f_ϑ with $\vartheta\neq\hat{\theta}$ in (10) is bounded by $\frac{\sum_{j\neq i}(1-W_n^j(\hat{\theta}_j))}{\alpha_1\prod_{j=1}^d W_n(\hat{\theta}_j)}$, and the weight placed on densities f_ϑ with $\vartheta_{-i}\neq\hat{\theta}_{-i}$ in (11) is bounded above by:

$$\frac{\sum_{s\in\Theta_i:s\neq\hat{\theta}_i}\alpha_2W_n^i(s)\sum_{j\neq i}(1-W_n^j(\hat{\theta}_j))}{\sum_{s\in\Theta_i:s\neq\hat{\theta}_i}\alpha_1W_n^i(s)\prod_{j\neq i}W^j(\hat{\theta}_j)}=\frac{\alpha_2\sum_{j\neq i}(1-W_n^j(\hat{\theta}_j))}{\alpha_1\prod_{j\neq i}W^j(\hat{\theta}_j)}.$$

For every $\eta > 0$, it follows that there is a δ satisfying $\frac{\alpha_2 d\delta}{\alpha_1(1-\delta)^d} < \eta$ such that when $W_n^j(\hat{\theta}_j) \geq 1-\delta$ for all j then the agent's beliefs in these ancillary states of the world which differ from t is bounded by η, conditioning either on $\theta_i = \hat{\theta}_i$ or $\theta_i \neq \hat{\theta}_i$. Suppose that (20) holds with $\delta > 0$ meeting this condition.

Now, application of (15) in Lemma 10 with $f_0 = f_{\theta^*}$, $f_1 = f_t$, $n_0 = 2$, $p_0 = 0$, and $p_1 = W_n^i(\hat{\theta}_i) \geq 1-\delta$ along with $\beta_1 = \beta^{-1}, \beta_2 = \beta$ implies that

$$\left| D_{\mathrm{KL}}\left(f_0 \,\middle|\middle|\, W_n(\cdot|\theta_i = \hat{\theta}_i)\right) - D_{\mathrm{KL}}\left(f_0 \,\middle|\middle|\, f_{\hat{\theta}}\right) \right| \leq K(\beta, \eta),$$

where $\lim_{\eta\to 0} K(\beta,\eta) = 0$. Similarly, by letting $f_1, \ldots, f_{d-1}$ be the elements of $G_i(t)$ and $n_0 = d$, one finds

$$\left| D_{\mathrm{KL}}\left(f_0 \,\middle|\middle|\, W_n(\cdot|\theta_i \neq \hat{\theta}_i)\right) - D_{\mathrm{KL}}\left(f_0 \,||\, f\right) \right| \leq K(\beta, \eta),$$

where $f \in \mathrm{co}\left(\{f_\theta : \theta \in G_i(t)\}\right)$. By (21), if one choose δ and hence η to be small enough so that $K(\beta,\eta) < \frac{L - D_{\mathrm{KL}}\left(f_{\theta^*} || f_{\hat{\theta}}\right)}{3} \equiv C$, the triangle inequality implies

$$D_{\mathrm{KL}}\left(f_0 \,\middle|\middle|\, W_n(\cdot|\theta_i \neq \hat{\theta}_i)\right) > C + D_{\mathrm{KL}}\left(f_0 \,\middle|\middle|\, W_n(\cdot|\theta_i = \hat{\theta}_i)\right).$$

where $C > 0$. Thus, by (12), (20) implies,

$$\mathbb{E}\left[\log \frac{w_n\left(Y_{n+1}\middle|\theta_i = \hat{\theta}_i\right)}{w_n\left(Y_{n+1}\middle|\theta_i \neq \hat{\theta}_i\right)}\right] = \int \left(\log \frac{w_n\left(Y_{n+1}\middle|\theta_i = \hat{\theta}_i\right)}{f_{\theta^*}} - \frac{\log w_n\left(Y_{n+1}\middle|\theta_i \neq \hat{\theta}_i\right)}{f_{\theta^*}}\right) f_{\theta^*}\,\mathrm{d}y > C. \tag{22}$$

Since the number of marginals to be considered is $d < \infty$, one can pick δ and C above small enough so that (22) holds for all i. Let $\Omega = Y^{\mathbb{Z}_+}$ where Ω is equipped with the product σ-algebra and product measure inherited from f_{θ^*} (since the observations Y_n are iid$\sim f_{\theta^*}$). For $1 \leq i \leq d$, define the following stochastic process $(X_n^i) : \Omega \to [0,1]$:

$$X_0^i = \log \frac{W_0^i(\hat{\theta}_i)}{1 - W_0^i(\hat{\theta}_i)}$$

$$X_{n+1}^i = \begin{cases} X_n^i & \text{if } \min_{\substack{1\le j\le d,\\ 0\le \ell\le n}} W_\ell^j(\hat{\theta}_j) < 1-\delta \\ X_n^i + \log \frac{w_n(Y_{n+1}|\theta_i = \hat{\theta}_i)}{w_n(Y_{n+1}|\theta_i \neq \hat{\theta}_i)} - C & \text{otherwise} \end{cases}$$

Then X_n^i is a submartingale with respect to the filtration $\mathscr{F}_n = \sigma\left(\left(W_\ell^j(\hat{\theta}_j)\right)_{\substack{1\le j\le d,\\ 0\le \ell\le n}}\right)$. Suppose that we pick $c \le \delta$ and $W_0^i(\hat{\theta}_i) \ge 1 - c$ for every i so that $X_0^i \ge \log \frac{1-c}{c}$. Then, let $\gamma = \log \frac{1-c}{c} - \log \frac{1-\delta}{\delta}$ and note that Azuma's inequality for submartingales implies that:

$$\begin{aligned} &\mathrm{P}_{\theta^*}\left(\bigcap_{i=1}^d \left\{\log \frac{W_1^i(\hat{\theta}_i)}{1-W_1^i(\hat{\theta}_i)} \ge \log \frac{1-\delta}{\delta} + \frac{C}{2}\right\}\right) \\ &\quad \ge 1 - \sum_{i=1}^d \mathrm{P}_{\theta^*}\left(\log \frac{W_1^i(\hat{\theta}_i)}{1-W_1^i(\hat{\theta}_i)} < \log \frac{W_0^i(\hat{\theta}_i)}{1-W_0^i(\hat{\theta}_i)} - \gamma + \frac{C}{2}\right) \\ &\quad \ge 1 - \sum_{i=1}^d \mathrm{P}_{\theta^*}\left(X_1^i - X_0^i < -\gamma - \frac{C}{2}\right) \\ &\quad \ge 1 - d\exp\left(\frac{-(\gamma + C/2)^2}{2\log\beta}\right). \end{aligned} \tag{23}$$

Let E_n denote the event $\bigcap_{\substack{1\le j\le d,\\ 0\le \ell\le n}} \left\{\log \frac{W_\ell^i(\hat{\theta}_i)}{1-W_\ell^i(\hat{\theta}_i)} \ge \log \frac{1-\delta}{\delta} + \frac{\ell C}{2}\right\}$. Then similarly,

$$\begin{aligned} \mathrm{P}_{\theta^*}(E_{n+1}|E_n) &= \mathrm{P}_{\theta^*}\left(\bigcap_{i=1}^d \left\{\log \frac{W_{n+1}^i(\hat{\theta}_i)}{1-W_{n+1}^i(\hat{\theta}_i)} \ge \log \frac{1-\delta}{\delta} + \frac{(n+1)C}{2}\right\} \middle| E_n\right) \\ &\ge 1 - \sum_{i=1}^d \mathrm{P}_{\theta^*}\left(\log \frac{W_{n+1}^i(\hat{\theta}_i)}{1-W_{n+1}^i(\hat{\theta}_i)} < \log \frac{W_0^i(\hat{\theta}_i)}{1-W_0^i(\hat{\theta}_i)} - \gamma + \frac{(n+1)C}{2} \middle| E_n\right) \\ &\ge 1 - \sum_{i=1}^d \mathrm{P}_{\theta^*}\left(X_{n+1}^i - X_0^i < -\gamma - \frac{(n+1)C}{2} \middle| E_n\right) \\ &\ge 1 - d\exp\left(-\frac{(\gamma + (n+1)C/2)^2}{2(n+1)\log\beta}\right)\mathrm{P}_{\theta^*}(E_n)^{-1}. \end{aligned}$$

Since $\mathrm{P}_{\theta^*}(E_n) = \mathrm{P}_{\theta^*}(E_1)\prod_{\ell=1}^{n-1}\mathrm{P}_{\theta^*}\left(E_{\ell+1}\middle|E_\ell\right)$ with $\mathrm{P}_{\theta^*}(E_1)$ as in (23), we obtain:

$$\mathrm{P}_{\theta^*}(E_n) \geq 1 - d\sum_{\ell=0}^{n-1}\exp\left(-\frac{(\gamma+(\ell+1)C/2)^2}{2(\ell+1)\log\beta}\right).$$

The sum on the right is majorized by $\sum_{\ell=0}^{\infty}\exp\left(-\frac{(\ell+1)C^2}{8\log\beta}\right)$, which converges, so by application of the DCT with respect to counting measure, $\lim_{\gamma\to\infty} d\sum_{\ell=0}^{\infty}\exp\left(-\frac{(\gamma+(\ell+1)C/2)^2}{2(\ell+1)\log\beta}\right) = 0$. It follows that γ, and thus c, can be chosen high enough so that $\mathrm{P}_{\theta^*}\left(\bigcap_{n=1}^{\infty}E_n\right) > 1-\varepsilon$, which concludes. □

C Simulation Results in the 2×2 Model

By Propositions 1 and Theorem 1, when $\theta = \sigma = H$ so that values of $Y = 2$ are very likely (or when $\theta = \sigma = L$ so that values of $Y = 0$ are very likely), then beliefs converge to the truth, regardless of initial priors. However, by Lemma 3 and Theorem 2, convergence is not guaranteed when $\theta \neq \sigma$. Accordingly, we consider simulations with $\theta = L$ and $\sigma = H$ (the order does not matter). We consider two sets of parameter robustness. First, we fix the priors P_0, Q_0 and see how the probability of divergence varies with the success probabilities (p_H, p_L, q_H, q_L). Second, we vary the priors, together with a cross section of success probabilities. For each set of parameters we solve one million simulations, running simulations until beliefs converge to zero or one, and calculate the fraction of simulations converging to the wrong value.

We set priors to $P_0 = 0.8$ and $Q_0 = 0.6$ (beliefs can possibly converge to the incorrect state), $q_H = 0.65$ and $q_L = 0.4$, and we do parameter sensitivity over a grid of (p_H, p_L) covering the full range. Figure 5 plots the frequency of simulation converging to the incorrect state. (Remember that the convergence probability is continuous in priors.) Not surprisingly, the convergence probability looks “multivariate Normal” as a function of the parameters.

When the states θ and σ are not so different, and $P_0 > Q_0$, then the simulations suggest that asymptotically $P_n \to 1$ and $Q_n \to 0$ with high probability. When the parameters differ, the frequency of convergence decreases (continuously) as they differ by more. When the states are far from symmetric, there is a robust region with positive probability of converging to the wrong

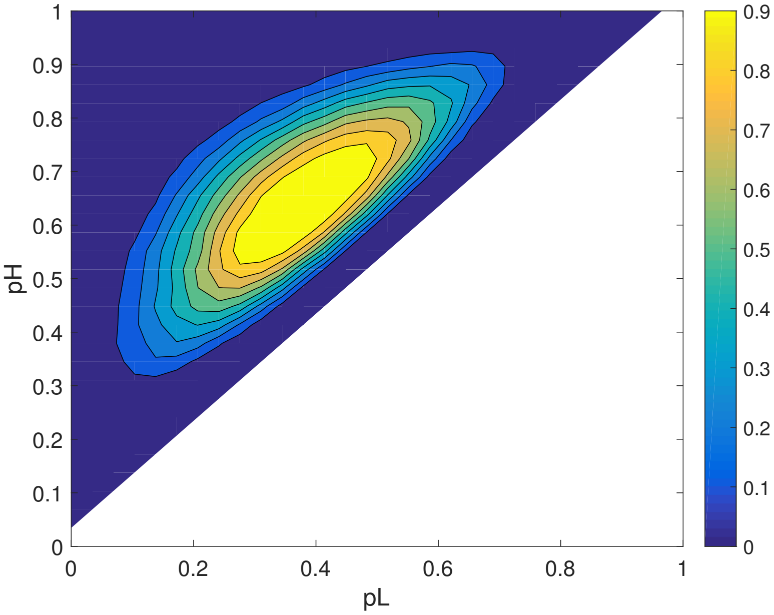
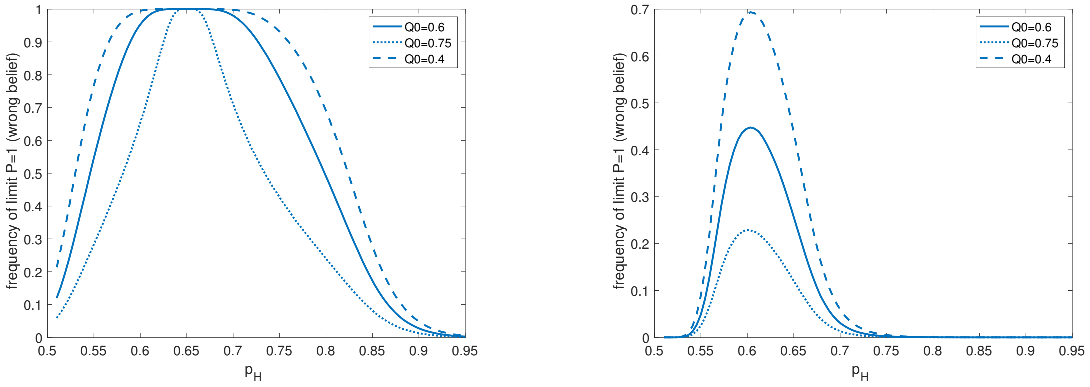


Figure 5: Probability of converging to wrong beliefs given prior $(P_0, Q_0) = (.8, .6)$, with $q_H = 0.65, q_L = 0.4$, varying p_H, p_L.

beliefs, but there need not be a set of parameters where asymptotic disagreement occurs with probability one.

To analyze a cross-section around symmetry, we set $p_L = 0.4$, and we do parameter sensitivity varying $p_H \in [0.5, 0.9]$ varying $Q_0 = 0.4, 0.6, 0.75$. The other parameters are as before. We then investigate a cross-section farther from symmetry. We run simulations varying p_H with $q_H = 0.4$ and $q_L = 0.2$, but this time with $p_L = 1 - p_H$. Figure 6a plots the frequency of simulations in which P_n converges to the incorrect value near symmetry. Figure 6b plots these results for when the states are quite asymmetric. The simulation evidence suggests that for a broad range of parameters $P_n \to 1$ with high probability, with divergence likely to occur the closer to symmetric are the states. Furthermore, the more different are P_0 and Q_0, the greater probability of $P_n \to 1$.

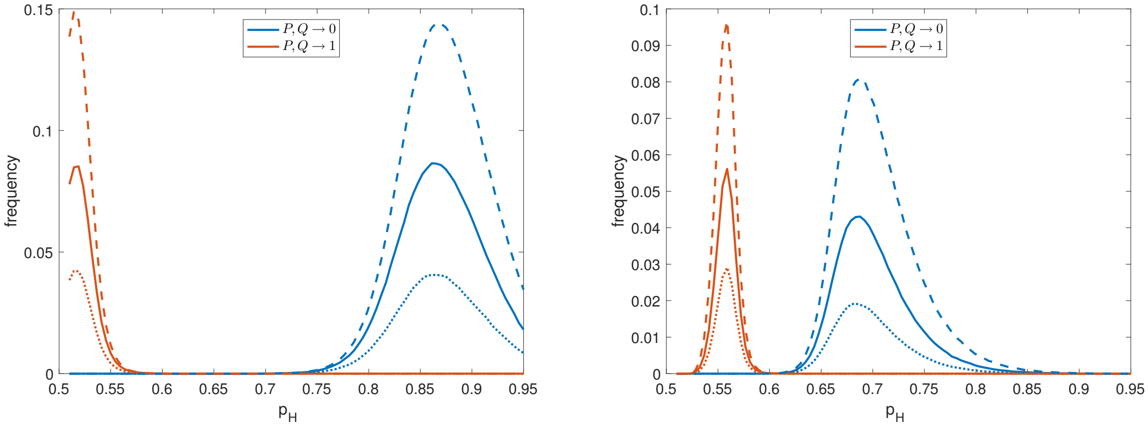
While the values of the aggregate states are complements, posteriors need not converge to complements (i.e., they could both converge to 1 or to 0). When states are not symmetric, posteriors can converge to values that are not even “symmetric” with the truth. This is surprising in the sense that beliefs converge to states that are observationally quite different. This result is important



(a) $P_0 = 0.8$, $q_H = 0.65$, $q_L = 0.4$, and $p_L = 0.4$. (b) $P_0 = 0.8$, $q_H = 0.4$, $q_L = 0.2$, $p_L = 1 - p_H$.

Figure 6: Probability of converging to wrong beliefs when starting with $P_0 > Q_0$.

because agents would not only disagree about the underlying values of θ, σ, but they would have quite different predictions for the distribution of Y_n. (One could argue that polarization truly refers to agents disagreeing about the value of $\theta + \sigma$, which is what would occur in this case.) The frequency of these convergences, of $P_n, Q_n \overset{\text{a.s.}}{\to} 1, 1$ and $P_n, Q_n \overset{\text{a.s.}}{\to} 0, 0$, are plotted in Figure 7, which shows the likelihood that both beliefs converge to zero for these parameters. Not surprisingly, the curve shifts up (more likely) when P_0 is more different from Q_0, down the closer they are.



(a) $q_H = 0.6, q_L = p_L = 0.4$. (b) $q_H = 0.4, q_L = 0.2, p_L = 1 - p_H$.

Figure 7: Probability of $P_n, Q_n \overset{\text{a.s.}}{\to} 1, 1$ and $P_n, Q_n \overset{\text{a.s.}}{\to} 0, 0$ when starting with $P_0 > Q_0$.