A Maximum Likelihood Approach to Combining Forecasts

Gilat Levy and Ronny Razin[1]

Abstract: We model an individual who wants to learn about a state of the world. The individual has a prior belief, and has data which consists of multiple forecasts about the state of the world. Our key assumption is that the decision maker identifies explanations that could have generated this data and among these focuses on the ones that maximise the likelihood of observing the data. The decision maker then bases her final prediction about the state on one of these maximum likelihood explanations. We show that in all the maximum likelihood explanations, moderate forecasts are just statistical derivatives of extreme ones. Therefore, the decision maker will base her final prediction only on the information conveyed in the relatively extreme forecasts. We show that this approach to combining forecasts leads to a unique prediction and a simple and dynamically consistent way of aggregating opinions.

1 Introduction

In many economic and political situations we find ourselves confronted with multiple opinions or forecasts about variables that are important for decision making. In financial markets we are often exposed to multiple forecasts about particular stocks or investment possibilities, when we buy a new laptop we might read multiple reviews either in news outlets or on social media platforms, and prior to election day we are exposed to multiple polls or to opinions espoused by friends and colleagues. These different pieces of information might be important for our decision making about investments, what we buy or who we vote for.

To aggregate forecasts, individuals need to take a view about how the forecasts are generated and how they are related to one another. Naturally, multiple forecasts may be correlated. This can arise when we aggregate advice from friends in a connected network, when election polling firms rely on the same data (as is often the case) or when forecasters in financial markets use similar data sources. Indeed, in the last two decades, online communication has introduced a more complicated web of information sources with potentially higher

[1]Levy: Department of Economics, LSE, g.levy1@lse.ac.uk. Razin: Department of Economics, LSE, r.razin@lse.ac.uk. We thank seminar participans in Northwestern, ESSET 2018, Bonn, the editor and two anonymous referees, as well as Xitong Hui for her research assistance. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 681579.

degrees of correlation across such forecasts. This complicates the problem of combining forecasts and puts into question traditional models that simply assume that individuals have a prior distribution about all relevant information structures generating these forecasts. One possible alternative assumption explored in the literature is that individuals use a simple heuristic, e.g. assume that forecasts are (conditionally) independent and aggregate them accordingly.[2]

In this paper, we take a novel approach to this problem: We assume that individuals aggregate forecasts by looking for the most likely explanation for what they had observed. That is, we consider a maximum likelihood theory of combining forecasts.[3] We model an individual who wants to learn about a state of the world (finite set of states). The individual has a prior belief and observes multiple forecasts about the state. These forecasts arrive sequentially, one in each period. We assume that the individual believes that each forecast was generated by a Bayesian forecaster but has no prior beliefs about the joint information structure generating the set of forecasts. Our key assumption is that she identifies joint information structures that had generated the data with the highest likelihood, and bases her final prediction on such explanations.

In our main result we characterise the set of maximum likelihood (ML) explanations and ML predictions that the individual adopts. We show that the ML prediction is unique and depends only on a small set of forecasts, not more than the number of states. In particular, the individual who uses a maximum likelihood explanation to combine forecasts will ignore forecasts that are relatively moderate and focus attention only on a set of extreme forecasts.

The intuition for this result follows from two simple observations. First, to increase the likelihood of her explanation, the individual "looks" for correlation between the sources of information that she thinks have generated the data. Correlated data sources are more likely to have generated a specific set of forecasts rather than independent sources. This implies that when interpreting the data through the lens of her explanation, she focuses attention on a small number of forecasts while ignoring others.

The second observation is that explaining extreme forecasts, that are more distant from the prior, will involve a lower maximum likelihood. To give an extreme example, consider the highest likelihood of observing a forecast that puts a probability close to one that the

[2]See for example DeMarzo et al (2003), Glaeser and Sunstein (2009), Levy and Razin (2015, 2018a) and Golub and Jackson (2012). Ellis and Piccione (2017) and Levy and Razin (2018b) consider decision makers who misperceive correlation. Spiegler (2016) analyses a model in which individuals mis-specify the true causal Bayesian network. Arieli et al (2019) use scoring rules and regret to evaluate forecast aggregation schemes.

[3]Such a procedure is in line with legal reasoning where plausible argumentation and balance of probabilities are suggested as normative approaches to the aggregation of information.

state is high. To explain this forecast we need to assume that the signals that generated it were very informative; thus the likelihood of observing this forecast is close to the prior probability that the state is indeed high. On the other hand, to explain a forecast that is close to the prior, one can assume an uninformative signal; an uninformative signal can generate a posterior which is close to the prior with a probability close to one. Therefore, explaining the extreme views puts an upper bound on the likelihood of any explanation of the entire data. As we show, the individual can achieve this bound. She explains extreme views and treats moderate forecasts as noisy derivatives of extreme forecasts.

The above result has several behavioural implications for the dynamic evolution of the ML prediction. First, ML predictions can exhibit stagnation. When more moderate forecasts are added to her data set, the decision maker does not change her prediction. Second, the prior plays an important role; specifically, the notion of extreme views is measured with respect to the prior. As we show, in the long term, limit predictions can depend on the prior even when individuals are exposed to rich and large amounts of information.

In our model the decision maker's misspecified view or narrative of the world is constantly changing with new observations. But does she have to change her "world view" drastically each time she observes a new forecast? We formulate a notion of time consistency across explanations under which older explanations are never discarded but continue to play a part in the new narrative the decision maker adopts. We show that the decision maker can always construct her ML explanations to be time-consistent. Intuitively, new forecasts are added to the previous explanation by either being correlated to previous extreme ones or the other way around. Thus, the decision maker's "world view" evolves to accommodate new forecasts without altering how previous ones are explained. Our formulation of time-consistency provides a novel normative criterion for such evolving misspecified models of the world.

A growing literature in economics studies how individuals combine forecasts in a non-Bayesian manner, with many papers focusing on correlation neglect. In contrast, in our analysis, decision makers entertain all possible correlation structures: The focus on maximising the likelihood implies that decision makers actively "look" for correlation.[4] While the correlation neglect assumption implies that individuals' beliefs become more extreme, in our model a different form of "extremism" arises, as individuals base their prediction only on relatively extreme forecasts. Our result complements other explanations for this form of extremism; for example, several papers show how extreme views are more likely to get into the consideration set of decision makers (either as those who espouse them have greater

[4]In this sense our approach is related to small-sample biases such as the "hot hand" fallacy (see Rabin and Vayanos 2010), where individuals tend to over-infer correlations in the data rather than neglect them.

incentive to do so, as in Osborne *et al* 2000 and Levy and Razin 2013, or as they are simpler to communicate, as in Levy and Razin 2012).

In statistics and economics the ML approach was first formalised by Fisher (1912) for the purpose of parameter estimation,[5] and by Robbins (1951), Good (1965) and Berger and Berliner (1986) for the purpose of forming posterior beliefs in the presence of ambiguity. Gilboa and Schmeidler (2003), Ortoleva (2012) and Suleymanov (2018) identify axioms that can rationalize maximum likelihood procedures while Gilboa and Schmeidler (2010) highlight the trade-off that can arise between maximising likelihood and simplicity. In Epstein and Schneider (2007), who model learning with ambiguity, a decision maker weeds out explanations that yield a likelihood less than a fraction times the maximum likelihood that can be achieved. Both Ortoleva (2012) and Epstein and Schneider (2007) use ML as a refinement criterion in extreme (zero or small) probability events, whereas we use ML as the main reasoning for decisions. Finally, Cherry and Salant (2018) suggest an equilibrium notion in which players best respond to a sample of their opponents' actions, where they make a statistical inference given this sample. One example they consider is for players to use ML to estimate the most likely parameter that has generated the sample.

2 The Model

A decision maker is forming a prediction about a state of the world $\omega \in \Omega$ where Ω is a finite set. She has a full support prior, $p \in \Delta(\Omega)$. At every period $t \in \{1, 2, \ldots\}$, the decision maker observes a forecast: A forecast $q_t \in \Delta(\Omega)$ is a (full support) probability distribution over Ω. Let $\mathbf{q}^t \in (\Delta(\Omega))^t$ denote the vector of forecasts up to period t. At any period t, the decision maker's observed history is $\mathbf{q}^t$.

At every t, the decision maker will combine these forecasts and the prior into a prediction, $\mu \in \Delta(\Omega)$, about the state. We assume that the decision maker thinks that at each period t, a single forecaster, with the same prior p, receives an informative signal and rationally derives q_t using Bayes rule.[6] To form a prediction, the decision maker will consider *explanations*, which are Bayesian models that are consistent with the observations $\mathbf{q}^t$ and the prior.

We now formally define what an explanation is. First, we define a joint information structure by a tuple $\mathcal{I} = (S, p, f(\mathbf{s}, \omega))$, such that $S = \times_{j=1}^{t} S_j$ is a finite set of signals and for $\mathbf{s} \in S$, $f(\mathbf{s}, \omega)$ is a joint probability distribution over signals and states $\omega \in \Omega$. Given $f(\mathbf{s}, \omega)$,

[5] Our analysis differs from the econometrics approach. In particular, our decision maker entertains a new information structure whenever new information arrives.

[6] We can extend the analysis to the case of non-common priors. The restrictions that forecasters use Bayes rule and that the priors are known imply that there is no over-fitting. If priors are both not known and not common then the maximum likelihood of observing a vector of forecasts is one.

when they are well defined, we have $f(\mathbf{s}|\omega)$, $f(\omega|\mathbf{s})$ and $f(\mathbf{s})$ for all $\omega \in \Omega$. Moreover, $f(\mathbf{s}, \omega)$ also implies the marginal distributions over the realisations of signal s_j, $f_j(s_j|\omega)$.

An explanation of the data implies that any forecast q_j in $\mathbf{q}^t$ was generated by a Bayesian forecaster who was exposed to a signal in S_j and knows the marginal distribution on signals in S_j, $f_j(s_j|\omega)$. Formally,

Definition 1 *An explanation of* $\mathbf{q}^t$ *is* $e = (\mathcal{I}^e, \mathbf{s}^e)$ *where* $\mathcal{I}^e$ *is a joint information structure and* $\mathbf{s}^e = (s_j^e)_{j=1}^t \in S^e$ *is a realisation of signals such that for all* $j \in \{1, 2, .., t\}$, $q_j(\omega) = \Pr^e(\omega|s_j^e) = \frac{p(\omega) f_j^e(s_j^e|\omega)}{\sum_{v\in\Omega} p(v) f_j^e(s_j^e|v)}$.

In other words, the decision maker perceives some information structure $\mathcal{I}$, and a particular realisation of signals $\mathbf{s}$, such that all the forecasts q_j can be rationalised by Bayes rule, assuming that each forecast q_j was based only on the signals generated by $f_j(s_j|\omega)$. Note that this formulation is general in the sense that it allows for the possibility that s_j contains information observed by other forecasters $j' \neq j$.

Let $\mathcal{E}(\mathbf{q}^t)$ be the set of explanations as defined above. We assume that the decision maker uses the ML criterion to select which explanation in $\mathcal{E}(\mathbf{q}^t)$ to adopt to set her prediction. The likelihood of observing $\mathbf{q}^t$ given an explanation $e \in \mathcal{E}(\mathbf{q}^t)$ is:

$$L(\mathbf{q}^t|e) = \sum_{v\in\Omega} p(v) f^e(\mathbf{s}^e|v)$$

Below we show that $\arg\max_{e\in\mathcal{E}(\mathbf{q}^t)} L(\mathbf{q}^t|e)$ exists, and thus we can define the ML prediction as follows:

Assumption 1 (ML prediction): At any period t, the decision maker's prediction satisfies

$$\mu^{ML}(\omega|\mathbf{q}^t) = \frac{p(\omega) f^{\hat{e}}(\mathbf{s}^{\hat{e}}|\omega)}{\sum_{v\in\Omega} p(v) f^{\hat{e}}(\mathbf{s}^{\hat{e}}|v)}, \text{ for some explanation } \hat{e} \in \arg\max_{e\in\mathcal{E}(\mathbf{q}^t)} L(\mathbf{q}^t|e)$$

Before proceeding to characterise ML explanations and predictions we introduce a useful Lemma that simplifies the analysis by reducing the set of explanations to those that have information structures with binary signals for each forecast:

Lemma 1: *Let* $e = (I^e, s^e)$ *be an explanation of* $\mathbf{q}^t$. *Then there exists an explanation* $e' = (I^{e'}, s^{e'})$ *of* $\mathbf{q}^t$ *with* $S^{e'} = \{s^*, s^{-*}\}^t$, *and* $L(\mathbf{q}^t|e) = L(\mathbf{q}^t|e')$.

Intuitively, the decision maker looks for an ex post rationalisation, which implies that she has in mind a vector of signals attained by all forecasters. As a result, for any information structure that rationalises the forecasts, we can construct an equivalent information structure which induces that vector of observed signals with the same probability, and bunches all other - unobserved - signals together. This new information structure produces the observed forecasts with the same likelihood, and moreover attains this with just two signals per forecaster.

3 The ML Prediction

In this section we characterise the ML prediction. We start with some helpful notation. Consider a history $\mathbf{q}^t$. For any forecast $j = 1, \ldots, t$, let $\alpha_j(\omega) = \frac{q_j(\omega)}{p(\omega)}$, let $\bar{\alpha}_j = \max_{v \in \Omega} \alpha_j(v)$ and $\omega^j \in \arg\max_{v \in \Omega} \alpha_j(v)$. Thus, for each forecast j, ω^j can be described as the state that becomes most surprising given the prior and $\bar{\alpha}_j$ is the magnitude of this largest surprise. Next, note that $\frac{\alpha_j(\omega^j)}{\bar{\alpha}_j} = 1$, whereas $\frac{\alpha_j(\omega)}{\bar{\alpha}_j} \leq 1$ for any $\omega \neq \omega^j$. For example, if $|\Omega| = 3$, and the prior is uniform, then for a forecast $\{0.7, 0.2, 0.1\}$ the vector $\frac{\alpha_j(\omega)}{\bar{\alpha}_j}$ is $\{1, \frac{0.2}{0.7}, \frac{0.1}{0.7}\}$. As can be seen in the proof, $\alpha_j(\omega)/\alpha_j(v)$ is inherently related to the constraint that each forecaster uses Bayes rule and specifically implies the probability that a forecast can be attained at state ω. Moreover, the most surprising state presents the upper bound of attaining a forecast, which is why the ratio $\alpha_j(\omega)/\bar{\alpha}_j$ will be important.

With this notation we can characterise the ML prediction of the decision maker:

Proposition 1: *Given* $\mathbf{q}^t$, *the ML prediction of the decision maker is unique and equals to, for any* $\omega \in \Omega$:

$$\mu^{ML}(\omega|\mathbf{q}^t) = \frac{p(\omega) \min_{j=1,\ldots,t}\{\frac{\alpha_j(\omega)}{\bar{\alpha}_j}\}}{\sum_{v \in \Omega} p(v) \min_{j=1,\ldots,t}\{\frac{\alpha_j(v)}{\bar{\alpha}_j}\}}$$

Proof of Proposition 1: Note that for any explanation e, for any $\omega \in \Omega$,

$$(1) f^e(\mathbf{s}^e|\omega) \leq \min_{j=1,\ldots,t} f_j^e(s_j^e|\omega).$$

As an explanation is an information structure that rationalises $\mathbf{q}^t$, we must have for any $\omega, \omega' \in \Omega$:

$$(2) \frac{q_j(\omega)}{q_j(\omega')} = \frac{p(\omega)}{p(\omega')} \frac{f_j^e(s_j^e|\omega)}{f_j^e(s_j^e|\omega')}.$$

Choose some state v. Equation (2) implies that by setting $f_j^e(s_j^e|v)$ at some level, we pin down all values $f_j^e(s_j^e|\omega)$, for any $\omega \in \Omega$.

Using (1) and (2), we can write the upper bound for the likelihood of any explanation as:

$$\begin{aligned} \sum_{\omega \in \Omega} p(\omega) f^e(\mathbf{s}^e|\omega) &\leq \sum_{\omega \in \Omega} p(\omega) \min_j \{f_j^e(s_j^e|\omega)\} \\ &= p(v)[\sum_{\omega \in \Omega} \min_j \{\frac{q_j(\omega)}{q_j(v)} f_j^e(s_j^e|v)\}] \end{aligned}$$

To find the explanation that maximises the likelihood of observing $\mathbf{q}^t$ (the left-hand-side above) we will maximise the right-hand-side and show that we can achieve a likelihood equal to the maximal upper bound.

First, to maximize the right-hand-side, note that the problem is increasing in $f_j^e(s_j^e|\upsilon)$ for any $j \leq t$. By (2), for any $\omega \in \Omega$, $f_j^e(s_j^e|\omega) = \frac{q_j(\omega)}{q_j(\upsilon)} \frac{p(\upsilon)}{p(\omega)} f_j^e(s_j^e|\upsilon) \leq 1 \Rightarrow f_j^e(s_j^e|\upsilon) \leq \frac{p(\omega)}{q_j(\omega)} \frac{q_j(\upsilon)}{p(\upsilon)}$ for any $\omega \in \Omega$. Recall however that ω^j maximizes $\frac{q_j(\omega)}{p(\omega)}$ and hence imposes the binding constraint on the upper bound on $f_j^e(s_j^e|\upsilon)$. Therefore, we set $f_j^e(s_j^e|\upsilon)$ at the upper bound,

$$f_j^e(s_j^e|\upsilon) = \frac{p(\omega^j)}{p(\upsilon)} \frac{q_j(\upsilon)}{q_j(\omega^j)} = \frac{\alpha_j(\upsilon)}{\bar{\alpha}_j}$$

We can then use this and (2) to derive for any other ω,

$$f_j^e(s_j^e|\omega) = \frac{\alpha_j(\omega)}{\bar{\alpha}_j}.$$

Second, we construct our ML explanation, e^*, to attain the upper bound. For any j, let the set of signals be $S_j = \{s^*, s^{-*}\}$. Then set for any ω,

$$f_j^{e^*}(s^*|\omega) = \frac{\alpha_j(\omega)}{\bar{\alpha}_j}, f_j^{e^*}(s^{-*}|\omega) = 1 - f_j^{e^*}(s^*|\omega).$$

By way of convention, order the signals so that s_j^* is "lower" than s_j^{-*}, and now set the cumulative marginal as $F_j^{e^*}(s^*|\omega) = f_j^{e^*}(s^*|\omega)$ and $F_j^{e^*}(s^{-*}|\omega) = 1$. As we have set the marginals, we can now set the joint distribution over signals. Note that given that we had integrated the Bayes rule constraint for each forecast, which imposes how the probability that a forecast is attained is related across the states, we can focus only on the conditional joint distributions (that is, conditional on each state ω). In particular, the joint unconditional distribution over signals will be determined first by the realisation of the state of the world and then by the distribution over signals conditional on this state.

Fix some $\omega \in \Omega$. To attain the upper bound on the joint distribution over the *observed* signals, designated to be the vector $\mathbf{s}^*$, set:

$$(3) F^{e^*}(\mathbf{s}^*|\omega) = f^{e^*}(\mathbf{s}^*|\omega) = \min_j \{f_j^{e^*}(s^*|\omega)\} = \min_j \{\frac{\alpha_j(\omega)}{\bar{\alpha}_j}\}$$

which implies that $F^{e^*}(\mathbf{s}^*|\omega) = \min_j F_j^{e^*}(s^*|\omega)$. To complete the explanation, continue by setting, for all $\mathbf{s}$,

$$F^{e^* s}(\mathbf{s}|\omega) = \min_j F_j^{e^*}(s_j|\omega)$$

The above is a proper distribution function as given a set of marginals (here, $F_j^{e^*}(s_j|\omega)$ for all $j \leq t$), there is always a joint distribution function that attains the upper Frechet bound, which is the one defined above.[7]

We have constructed an explanation that consists of a set of conditional joint distribution functions and that achieves the upper bound. Note moreover that any explanation that

[7] See Joe (1977).

achieves it, must satisfy $f^{e^*}(\mathbf{s}^*|\omega) = \min_j\{\frac{\alpha_j(\omega)}{\bar{\alpha}_j}\}$ for some observed vector of signals $\mathbf{s}^*$. As a result, the ML prediction is unique and satisfies:

$$\mu^{ML}(\omega|\mathbf{q}^t) = \frac{p(\omega)f^{e^*}(\mathbf{s}^*|\omega)}{\sum_{v\in\Omega} p(v)f^{e^*}(\mathbf{s}^*|v)} = \frac{p(\omega)\min_{j=1,\ldots,t}\{\frac{\alpha_j(\omega)}{\bar{\alpha}_j}\}}{\sum_{v\in\Omega} p(v)\min_{j=1,\ldots,t}\{\frac{\alpha_j(v)}{\bar{\alpha}_j}\}}$$

which is the expression in the Proposition.■

From the above it is clear that only a small set of forecasts would matter, those that minimize $\frac{\alpha_j(\omega)}{\bar{\alpha}_j}$ for some ω. Thus, at most $|\Omega|$ forecasts would be relevant for the final prediction, whereas other forecasts can be ignored. Moreover, the order in which these forecasts arrive before or at time t will not alter the prediction at time t. Note also the importance of the prior: The prior determines $\frac{\alpha_j(\omega)}{\bar{\alpha}_j}$ and hence affects which forecasts can be ignored and how those that are not ignored are combined into a prediction (we return to the role of the prior when we consider the limit predictions in Section 4.2).

In Section 3.1 below, we illustrate how the above result translates to a simple prediction rule in the binary state space, and also construct an example of a ML explanation for $t = 2$. The ML explanation involves large degrees of correlation and specifically it is moderate forecasts that are correlated to extreme ones, and can therefore be ignored in the ML prediction. We show this feature more generally in section 3.2.

3.1 Binary states: ML prediction and explanation

We now illustrate Proposition 1 (as well as its proof) in the binary state space, $\Omega = \{0, 1\}$. Assume without loss of generality a prior of $p(1) = p > \frac{1}{2}$. To simplify, with some abuse of notation, let q_j denote the probability that the state is 1 according to the forecast in period j. Note that in the binary space, $\omega^j = 1$ if $q > p$ and $\omega^j = 0$ if $q < p$.

Consider for the sake of exposition $t = 2$, and first assume that $q_1 > q_2 > p$. Thus, both forecasts "support" state 1 compared with the prior (so that $\omega^1 = \omega^2 = 1$).

Suppose, in line with Lemma 1, that each forecaster can receive two signal realisations, s^* and s^{-*}, which, given Bayesian updating, must satisfy then, for some $\beta_j \in (0, 1]$:

$$\begin{aligned} f^e(s_1 &= s^*|\omega = 1) = \beta_1,\ f^e(s_1 = s^*|\omega = 0) = \beta_1\frac{p(1-q_1)}{q_1(1-p)} = \beta_1\frac{\alpha_1(0)}{\bar{\alpha}_1}, \\ f^e(s_2 &= s^*|\omega = 1) = \beta_2,\ f^e(s_2 = s^*|\omega = 0) = \beta_2\frac{p(1-q_2)}{q_2(1-p)} = \beta_2\frac{\alpha_2(0)}{\bar{\alpha}_2} \end{aligned}$$

Maximising β_1 and β_2 will increase the marginal probability of attaining the s^* signals for both forecasters, as $q_j > p$, and so we can set $\beta_j = 1$. We continue then by setting the joint probability over (s^*, s^*) :

$$f^e(s_1 = s^*, s_2 = s^*|\omega = 1) = 1 = f^e(s_1 = s^*|\omega = 1)$$
$$f^e(s_1 = s^*, s_2 = s^*|\omega = 0) = \min_j \frac{\alpha_j(0)}{\bar{\alpha}_j} = \frac{p(1-q_1)}{q_1(1-p)} = f^e(s_1 = s^*|\omega = 0).$$

Forecast 2 will be completely ignored as conditional on $s_1 = s^*$, the joint information structure constructed above necessitates $s_2 = s^*$ as well, and so the "observed" signal of the second forecaster is fully correlated with that of the first forecaster in each state. This readily implies the full ML joint information structure, depicted below, where each cell in each matrix denotes $f^e(s_1, s_2|\omega)$:

$\omega = 1$	$s_2 = s^*$	$s_2 = s^{-*}$	$\omega = 0$	$s_2 = s^*$	$s_2 = s^{-*}$
$s_1 = s^*$	1	0	$s_1 = s^*$	$\frac{p(1-q_1)}{q_1(1-p)}$	0
$s_1 = s^{-*}$	0	0	$s_1 = s^{-*}$	$\frac{p(1-q_2)}{q_2(1-p)} - \frac{p(1-q_1)}{q_1(1-p)}$	$1 - \frac{p(1-q_2)}{q_2(1-p)}$

As a result the ML prediction satisfies:

$$\mu = \frac{p}{p + (1-p)\frac{p(1-q_1)}{q_1(1-p)}} = q_1,$$

implying that the more moderate forecast, q_2, is completely ignored.

Assume instead that $q_2 < p < q_1$. To increase the likelihood of attaining q_2, a forecast indicating that the state is more likely to be 0 compared with the prior, we need that in state 1 the second forecaster does not always observe s^*. Specifically, we would now have:

$$f^e(s_1 = s^*|\omega = 1) = 1,\ f^e(s_1 = s^*|\omega = 0) = \frac{p(1-q_1)}{q_1(1-p)},$$
$$f^e(s_2 = s^*|\omega = 1) = \frac{(1-p)q_2}{p(1-q_2)},\ f^e(s_2 = s^*|\omega = 0) = 1.$$

This implies the following joint ML information structure, in which the "observed" signal of forecaster 1 is a sufficient statistic for the "observed" signal of forecaster 2 only in state 0, and the other way around in state 1:

$\omega = 1$	$s_2 = s^*$	$s_2 = s^{-*}$	$\omega = 0$	$s_2 = s^*$	$s_2 = s^{-*}$
$s_1 = s^*$	$\frac{(1-p)q_2}{p(1-q_2)}$	$1 - \frac{(1-p)q_2}{p(1-q_2)}$	$s_1 = s^*$	$\frac{p(1-q_1)}{q_1(1-p)}$	0
$s_1 = s^{-*}$	0	0	$s_1 = s^{-*}$	$1 - \frac{p(1-q_1)}{q_1(1-p)}$	0

As a result, the ML prediction satisfies now:

$$\mu = \frac{p\frac{(1-p)q_2}{p(1-q_2)}}{p\frac{(1-p)q_2}{p(1-q_2)} + (1-p)\frac{p(1-q_1)}{q_1(1-p)}} = \frac{q_2 q_1}{q_2 q_1 + \frac{p}{1-p}(1-q_1)(1-q_2)}$$

Corollary 1, derived from Proposition 1, shows how the above extends to any t:

Corollary 1: *Suppose that* $\Omega = \{0,1\}$. *Let* $q_t^{\max} = \max_{j\leq t} q_j$ *and let* $q_t^{\min} = \min_{j\leq t} q_j$. *For any* $\mathbf{q}^t$, *the ML prediction is:*

$$\mu^{ML}(1|\mathbf{q}^t) = \begin{cases} q_t^{\max} \text{ if } q_t^{\min} \geqslant p \\ q_t^{\min} \text{ if } q_t^{\max} \leqslant p \\ \frac{q_t^{\min} q_t^{\max}}{q_t^{\min} q_t^{\max} + \frac{p}{(1-p)}(1-q_t^{\min})(1-q_t^{\max})} \text{ otherwise} \end{cases}$$

The implication of Proposition 1 to the binary state space is that at most two forecasts, the most extreme ones (on each side of the prior), will matter for the ML prediction. All other - more moderate - forecasts will be ignored.

3.2 Ignoring moderate forecasts

In this section we characterise the set of forecasts that will be ignored in the ML prediction. Formally, we say that forecast q_{t+1} is *ignored* if $\mu^{ML}(\omega|\mathbf{q}^{t+1}) = \mu^{ML}(\omega|\mathbf{q}^t)$.

Remember that for any forecast q_j, $\alpha_j(\omega)$ is the ratio of the probability of this state under the forecast divided by the prior probability. We have also introduced the notation of $\bar{\alpha}_j = \max_{v\in\Omega} \alpha_j(v)$ and $\omega^j \in \arg\max_{v\in\Omega} \alpha_j(v)$. Thus, for each forecast q_j, ω^j can be described as the state that becomes most surprising given the prior and $\bar{\alpha}_j$ is the magnitude of this largest surprise.

Going back to the proof of Proposition 1, note that Bayesian updating bounds the probability that a forecast can be sent at each state ω. A forecast q_j, which is almost degenerate on state ω^j - hence, extreme - can be sent in any other state ω with only a very small probability as we would have $\frac{\alpha_j(\omega)}{\bar{\alpha}_j} \approx 0$; otherwise it could not indicate to a rational forecaster that the state is most likely to be ω^j. Thus, this forecast would bound the joint probability of obtaining all forecasts in all states $\omega \neq \omega^j$ and would therefore matter for the ML prediction. On the other hand, a moderate forecast which agrees with the prior, $q_j(\omega) = p(\omega)$, can be sent in each state ω with probability $\frac{\alpha(\omega)}{\bar{\alpha}_j} = 1$. It will then never impose a bound on the joint probability distribution and would not affect the ML prediction. We now characterise the set of moderate forecasts, such as above, that will be ignored in the final aggregation of forecasts. Let:

$$I(\mathbf{q}^t) = \left\{ q_{t+1} \in \Delta(\Omega) | \begin{array}{c} \forall v \in \Omega, \frac{\alpha_{t+1}(v)}{\bar{\alpha}_{t+1}} \geq \min_{j\leq t} \frac{\alpha_j(v)}{\bar{\alpha}_j} \\ \text{and } \exists \omega \in \Omega \text{ such that } \frac{\alpha_{t+1}(\omega)}{\bar{\alpha}_{t+1}} = \min_{j\leq t} \frac{\alpha_j(\omega)}{\bar{\alpha}_j} \end{array} \right\}$$

The set $I(\mathbf{q}^t)$ defines the boundary of the set of all forecasts that the decision maker will be able to ignore at period $t+1$. This boundary set contains forecasts that minimise $\frac{\alpha_j(\omega)}{\bar{\alpha}_j}$

for some state ω, for $j \leq t$, and hence contains forecasts that have not been ignored up to and including period t. To gain intuition, let us consider the set $I(\mathbf{q}^t)$ in the case of binary states. Suppose that $\min_{j\leq t} q_j < p < \max_{j\leq t} q_j$. Assume without loss of generality that $\max_{j\leq t} q_j > 1 - \min_{j\leq t} q_j$ so that $\bar{\alpha}_j = \max_{j\leq t} q_j$. In this case we have that

$$\min_{j\leq t} \frac{\alpha_j(1)}{\bar{\alpha}_j} = \frac{\min_{j\leq t} q_j}{\max_{j\leq t} q_j} \text{ and } \min_{j\leq t} \frac{\alpha_j(0)}{\bar{\alpha}_j} = \frac{1 - \max_{j\leq t} q_j}{\max_{j\leq t} q_j}.$$

Therefore, if $q_{t+1} = \max_{j\leq t} q_j$ then

$$\frac{\alpha_{t+1}(1)}{\bar{\alpha}_{t+1}} = 1 > \frac{\min_{j\leq t} q_j}{\max_{j\leq t} q_j} \text{ and } \frac{\alpha_{t+1}(0)}{\bar{\alpha}_{t+1}} = \frac{1 - \max_{j\leq t} q_j}{\max_{j\leq t} q_j} = \min_{j\leq t} \frac{\alpha_j(0)}{\bar{\alpha}_j}.$$

Similarly when $q_{t+1} = \min_{j\leq t} q_j$ we have that

$$\frac{\alpha_{t+1}(1)}{\bar{\alpha}_{t+1}} = \frac{\min_{j\leq t} q_j}{\max_{j\leq t} q_j} = \min_{j\leq t} \frac{\alpha_j(1)}{\bar{\alpha}_j} \text{ and } \frac{\alpha_{t+1}(0)}{\bar{\alpha}_{t+1}} = \frac{1 - \min_{j\leq t} q_j}{\max_{j\leq t} q_j} > \frac{1 - \max_{j\leq t} q_j}{\max_{j\leq t} q_j} = \min_{j\leq t} \frac{\alpha_j(0)}{\bar{\alpha}_j}.$$

For any $q_{t+1} \in (\min_{j\leq t} q_j, \max_{j\leq t} q_j)$ both inequalities will be strict and for any $q_{t+1} \notin [\min_{j\leq t} q_j, \max_{j\leq t} q_j]$ one of the inequalities will hold in the opposite way. As a result, in this case we have $I(\mathbf{q}^t) = \{\min_{j\leq t} q_j, \max_{j\leq t} q_j\}$.

Using the definition of the set $I(\mathbf{q}^t)$ we show for the general state space:

Lemma 2: *For any $q \in I(\mathbf{q}^t)$ and $\beta \in [0,1]$, the forecast $q_{t+1} = \beta p + (1-\beta)q$ is ignored.*

If we define moderation by the distance to the prior, Lemma 2 implies that all forecasts that are more moderate (weakly) than those in $I(\mathbf{q}^t)$ will be ignored. In the case of binary states, when $\min_{j\leq t} q_j < p < \max_{j\leq t} q_j$, we saw above that $I(\mathbf{q}^t) = \{\min_{j\leq t} q_j, \max_{j\leq t} q_j\}$. Thus, all more moderate forecasts on $[\min_{j\leq t} q_j, p]$ and on $[p, \max_{j\leq t} q_j]$ will be ignored. To see another example, consider a tertiary state space, $\Omega = \{0,1,2\}$. With equal prior, the forecast $\{0.7, 0.2, 0.1\}$ allows us to ignore all forecasts for whom the vector $\frac{\alpha_j(\omega)}{\bar{\alpha}_j}$ is at least, element by element, $\{1, \frac{0.2}{0.7}, \frac{0.1}{0.7}\}$. The set of such ignored forecasts is depicted in Figure 1. As can be seen, it is all those that have the same mode, at $\omega = 0$, and lie on the line that connects each point on the closure to the prior.

Figure 2 shows the set of forecasts that will be ignored when $\{0.7, 0.2, 0.1\}$ is observed, and the set of forecasts that will be ignored when $\{0.2, 0.1, 0.7\}$ is observed. These are the two light shaded areas in the figure. The set of forecasts that will be ignored when both $\{0.7, 0.2, 0.1\}$ and $\{0.2, 0.1, 0.7\}$ are observed includes more forecasts than just the union of the above two sets. The additional forecasts that will be ignored when both $\{0.7, 0.2, 0.1\}$ and $\{0.2, 0.1, 0.7\}$ are observed is given by the darker shaded area in the figure. Thus the total set of forecasts that will be ignored when both $\{0.7, 0.2, 0.1\}$ and $\{0.2, 0.1, 0.7\}$ are observed includes the union of the three shaded areas depicted in the figure.

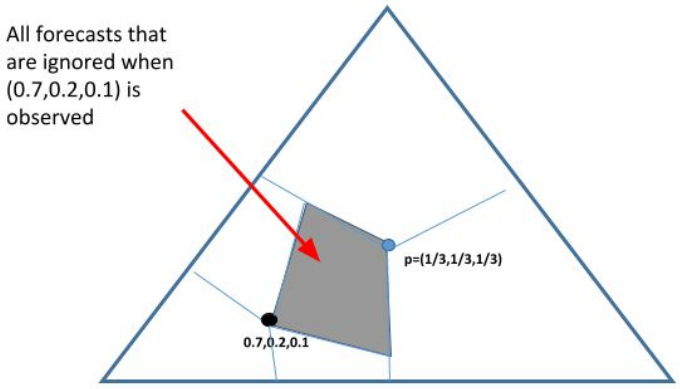

Figure 1

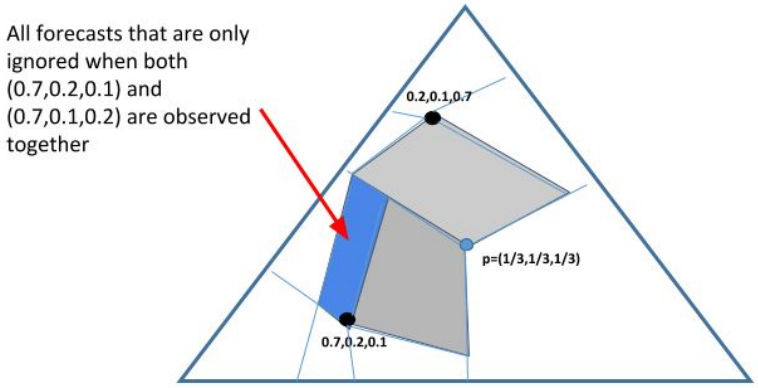

Figure 2

3.3 Time consistency of ML explanations

The ML decision maker has potentially a wrong model in mind, and with every new observation she adopts a different "world view" or narrative to interpret the accumulated data she has observed. However, if the procedure calls for a completely new model each period, we might question the sensibility of this method of aggregating data. Moreover, such a procedure would imply higher computation costs, as each period a new model would need to be calculated.

Below we formalise a notion of consistency between explanations and show that the ML explanation can evolve in a time-consistent manner. Let e be an explanation of $\mathbf{q}^t$ and e' an explanation of $\mathbf{q}^{t+1}$. We say that e' is time consistent with e if $S^{e'} = S^e \times S_{t+1}$ for some finite set S_{t+1} and for any $\mathbf{s} \in S^e, \omega \in \Omega$,

$$f^e(\mathbf{s}|\omega) = \sum_{(\mathbf{s},s_{t+1})\in S^{e'}} f^{e'}(\mathbf{s}, s_{t+1}|\omega)$$

In words, $f^e(\mathbf{s}|\omega)$, the probability of observing $\mathbf{s}$ in explanation e is equal to the expected probability of observing signal realisations $\mathbf{s}$ in the first t periods under explanation e'. This implies that the explanation e for the first t observations is the projection of e' to these t observation. As a result if there is time consistency between e and e' we can say that when moving from period t to period $t+1$, the decision maker keeps, in a statistical sense, her explanation for the first t observations.

We can then show:

Proposition 2: *For any explanation of* e *of* $\mathbf{q}^t$, for *any observation* q_{t+1}, *there exists an ML explanation* e' *of* $(\mathbf{q}^t, q_{t+1})$ which is time consistent *with* e.

Intuitively, for any state ω, there are two cases: Either the new forecast is ignored, which means that the probability of observing it is higher than observing a previous one and so we can fully correlate it with the previous vector of forecasts and peg the remaining probability of observing the new forecast on to the other, unobserved, realisations of the first t forecasts. Or, old forecasts are ignored so we fully correlate old forecasts to the new one and the remaining probability of observing the old forecasts is pegged on to other (unobserved) realisations of the $t+1$ forecast. As a result, the decision maker does not need to come up with a completely new ML explanation each time and can be consistent in how she explains her data over time.

To get more intuition, recall the joint information structure described in the case of binary states, when $q_1 > q_2 > p$, which, for the "observed" signals, satisfied:

$$f^e(s_1 = s^*, s_2 = s^*|\omega = 1) = 1, f^e(s_1 = s^*, s_2 = s^*|\omega = 0) = \frac{p(1-q_1)}{q_1(1-p)}.$$

Assume that we now observe some q_3 where specifically we have $q_3 < p$, for which the ML information structure will demand a marginal probability of:

$$f^e(s_3 = s^*|\omega = 1) = \frac{(1-p)q_3}{p(1-q_3)}, f^e(s_3 = s^{-*}|\omega = 0) = 1$$

Consider the following joint ML structure, where we combine the above marginal of the first two signals, s_1 and s_2, with that of s_3. First, we describe $f^e(s_1, s_2, s_3 = s^*|1)$ and $f^e(s_1, s_2, s_3 = s^{-*}|1)$:

$\omega = 1,\ s_3 = s^*$	$s_2 = s^*$	$s_2 = s^{-*}$		$\omega = 1,\ s_3 = s^{-*}$	$s_2 = s^*$	$s_2 = s^{-*}$
$s_1 = s^*$	$\frac{(1-p)q_3}{p(1-q_3)}$	0	,	$s_1 = s^*$	$1 - \frac{(1-p)q_3}{p(1-q_3)}$	0
$s_1 = s^{-*}$	0	0		$s_1 = s^{-*}$	0	0

We now depict $f^e(s_1, s_2, s_3 = s^*|0)$ and $f^e(s_1, s_2, s_3 = s^{-*}|0)$:

$\omega = 0,\ s_3 = s^*$	$s_2 = s^*$	$s_2 = s^{-*}$		$\omega = 0,\ s_3 = s^{-*}$	$s_2 = s^*$	$s_2 = s^{-*}$
$s_1 = s^*$	$\frac{p(1-q_1)}{q_1(1-p)}$	$\frac{p(1-q_2)}{q_2(1-p)} - \frac{p(1-q_1)}{q_1(1-p)}$	,	$s_1 = s^*$	0	0
$s_1 = s^{-*}$	0	$1 - \frac{p(1-q_2)}{q_2(1-p)}$		$s_1 = s^{-*}$	0	0

The marginal distribution over s_1, s_2 derived from the above is exactly as it is for the two signals generating $q_1 > q_2 > p$, as described in Section 3.1. Moreover, following from Proposition 1, the above is an ML information structure.

The forecast q_3 cannot be ignored and the signal generating it, $s_3 = s^*$, has the minimum probability of being attained in state 1 compared with the probability of attaining $s_1 = s_2 = s^*$. As a result, these signals are fully correlated to $s_3 = s^*$ in state 1 and the remaining probability of attaining $s_1 = s_2 = s^*$ in this state is pegged on to the (non) observation of $s_3 = s^{-*}$. In state 0, the joint probability of attaining $s_1 = s_2 = s^*$ is lower than that of $s_3 = s^*$, which is one. As a result, $s_3 = s^*$ arises whenever $s_1 = s_2 = s^*$ arises but also arises for other configurations of these signals. The proof in the Appendix uses similar arguments to construct a time-consistent joint distribution between the marginal of the first t signals and the additional signal arising in the $t + 1th$ period.

4 Extensions

We consider two extensions. In the first one we analyse a model in which the decision maker does not observe forecasts but only the history of her past ML predictions. The second extension looks at limit predictions, and highlights the importance of the prior belief in the ML process. To simplify the exposition, we henceforth focus on a binary state space $\Omega = \{0, 1\}$. Thus a forecast q_j simply denotes the probability that the state is 1.

4.1 Source amnesia

We have so far assumed that the decision maker remembers the history of forecasts she has seen. We now consider an extension to a decision maker that does not remember forecasts but does remember her past predictions. This form of memory is termed *source amnesia* in the psychological literature.[8]

Specifically, in our previous model, at time t, the decision maker observes $\mathbf{q}^t = (q_1, q_2, \ldots, q_t)$. We now assume that at time t, a decision maker observes q_t, and $\boldsymbol{\mu}^{t-1} = (\mu_1, \mu_2, \ldots, \mu_{t-1})$, where μ_j is the ML prediction at period j.

Assume that the prior is uniform, and suppose without loss of generality that $q_1 > \frac{1}{2}$. Let us consider Period 3, where the decision maker observes q_3, μ_2 and $\mu_1 = q_1$. She knows that for whichever q_2 she imagines, it has to be that $\mu_2 = \mu^{ML}(.|\mathbf{q}^2)$. By Corollary 1, if $\mu_2 \geq q_1$, then $\mu_2 = q_2$. If $\mu_2 = q_1$, then $q_1 > q_2 > \frac{1}{2}$ and it is not important for her to know the exact value of q_2. If $\mu_2 < q_1$, then $q_2 = \frac{\mu^2(1-q_1)}{q_1+\mu^2(1-2q_1)}$. Thus, in Period 3 she either knows q_2 or knows that it should be ignored. It follows then that at Period t, for each $j < t$, she either knows q_j or knows that it could be ignored. We therefore have:

Proposition 3: *A decision maker that has source amnesia will have the same ML prediction as a decision maker that remembers the history of forecasts.*

Remark 1 (*Social learning*): Social learning is typically modelled as an environment in which an individual at period t observes the sequence of the actions or predictions of her predecessors, as well as her own forecast. Thus as above, adopting social learning (with observed beliefs) to our setup, she will observe $\boldsymbol{\mu}^{t-1} = \{\mu_1, \mu_2, \ldots, \mu_{t-1}\}$, as well as q_t. If we assume that every decision maker believes that every one before her also uses the ML procedure, and that this is common knowledge, then the problem becomes identical to the single agent problem. Given this, all the results above hold. Specifically, when all previous predictions are observed, then the sequence of predictions will be identical to the one that arises if all forecasts were shared instead.

4.2 The Persistence of the prior

In this section we explore the convergence properties of the ML predictions. The key observation that we highlight is the importance of the prior in the limit ML predictions. Specifically, we show that even when exposed to a large data set, the individual's limit belief might still depend on her initial prior.

[8] See for example Schacter et al (1984).

To focus our discussion, assume a true information generating process by which signals are drawn in an iid manner each period. In particular, assume that the information structure induces a distribution F over forecasts in $(0,1)$, with $a = \inf(SupportF)$ and $b = \sup(SupportF)$. Assume that the prior p satisfies $p \in (a,b)$ and $0 \leq a < b \leq 1$, so that the posteriors are informative. When the individual observes forecasts drawn from this process, where would her predictions converge to?

Proposition 4:

(i) Suppose that $a > 0$ and $b < 1$. The prediction of the ML decision maker satisfies for any $\varepsilon > 0$

$$\lim_{t\to\infty} \Pr[\mu^{ML}(1|\mathbf{q}^t) \in (\frac{1}{1+\frac{p}{1-p}\frac{(1-a)}{b}\frac{(1-b)}{a}} - \varepsilon, \frac{1}{1+\frac{p}{1-p}\frac{(1-a)}{b}\frac{(1-b)}{a}} + \varepsilon)] = 1.$$

(ii) Suppose that $a = 0$ and $b = 1$. There exist distributions F for which for any $\varepsilon, \delta > 0$,

$$\lim_{t\to\infty} \Pr[\mu^{ML}(1|\mathbf{q}^t) \in (p-\varepsilon, p+\varepsilon)] > \delta.$$

The ML decision maker will update just based on the extreme forecasts she has observed along the sequence, on each side of the prior. To see (i), note that in the limit, by the law of large numbers, these extreme forecasts will converge to be a and b. Therefore, by Proposition 1, the ML prediction will converge to the expression in the Proposition. As $0 < \frac{(1-a)}{b}\frac{(1-b)}{a} < \infty$, the limit prediction still depends on the prior.

But what happens when $[a,b] = [0,1]$, that is when we have a rich signal structure that can induce posteriors of unlimited accuracy? For case (ii) we need to look at the limit statistical properties of $\tilde{q}_t^{\max}$ and $\tilde{q}_t^{\min}$ as they will determine whether the two posteriors will converge or not. Specifically, consider a distribution over posteriors conditional on the state $\omega = 1$. Given that the state is 1, the probability of observing a posterior close to 0 must be close to zero. This is because such a posterior is based on Bayesian updating and so for a forecaster to believe that the state is likely to be 0, it must be that this posterior is sent in state 1 only with a small probability. This implies that the distribution over posteriors given $\omega = 1$ must have a "thin tail" around zero. But this distribution can also have a similar thin tail around posteriors that are close to 1. Extreme value theory tells us that we can construct distributions of $\tilde{q}_t^{\max}$ and $\tilde{q}_t^{\min}$ to accord, in the limit, with the Gumbel distribution.[9] This allows us to construct information structures for which there is a strictly positive probability that the limit of the ratio $\frac{1-\tilde{q}_t^{\max}}{\tilde{q}_t^{\min}}$ is bounded and equals one, resulting in the ML belief converging to the prior p.

Remark 2 *(Divergence)*: Consider now two ML decision makers with different priors, where each assumes that the forecasts she observes are generated according to an information

[9] For example see Fisher and Tippet (1928).

structure incorporating her own prior. The above result implies that there is a strictly positive probability of divergence between the two ML predictions they will generate, even after being exposed to the same rich and large set of observations. This differs from the convergence properties of Bayesian decision makers with heterogeneous priors. For example, in a recent paper Acemoglu *et al* (2016) show that decision makers with heterogeneous priors, who agree about the information process generating their signals, will converge to have the same beliefs.[10]

5 Conclusion

There is some recent empirical and experimental evidence showing that decision makers tend to neglect correlation in some environments (e.g., Ortoleva and Snowberg (2015), Enke and Zimmerman (2019), Kallir and Sonsino (2009) and Eyster and Weiszacker (2011)) while in other contexts they overestimate correlation: Consistent with the maximum likelihood approach we assumed in this paper, De Filippis *et al* (2017) and Hossain and Okui (2018), derive results in which subjects sometimes overestimate the level of correlation.[11] It would be interesting to understand empirically when it is more likely for decision makers to become aware of, or alternatively excessively consider, correlation in their observed forecasts. One possibility is that observing many repeated forecasts will increase the suspicion of the decision maker that these are correlated. One interesting example for such reasoning is the Talmudic Sanhedrin court law that requires that if judges are unanimous in conviction, the defendant should be set free, while if only a majority convict, this majority verdict pertains. Glatt (2013) offers a maximum likelihood rationalisation of this rule; unanimity among many judges most likely is a result of strong correlation between the judges, and therefore demands caution. Gunn *et al* (2016) discuss this interpretation also in other legal scenarios.

6 Appendix

Proof of Lemma 1: Assume that $e = (I^e, \mathbf{s}^e)$ is an explanation of $\mathbf{q}^t$. We will construct a new explanation of $\mathbf{q}^t$ with $S_j = \{s^*, s^{-*}\}$ for any j, which maintains the same likelihood of observing $\mathbf{q}^t$ as e does.

[10] See section 3.1 in their paper. The paper uses this result as a benchmark to show, in the main part of the paper, that small disagreement about the information generating process can imply large disagreements about where the process converges to.

[11] Specifically, De Filippis *et al* (2017) consider a class of updating rules that generalize the ML updating one, and their experimental evidence supports the rule with a biased cirtical value (rather than an unbiased one).

Specifically, construct the new explanation $e' = (I^{e'}, \mathbf{s}^{e'})$ as follows. Let $\mathbf{s}^{e'} = (s^*, s^*, \ldots, s^*)$. Let, $\forall \omega \in \Omega$:

$$(*) f_j^{e'}(s^*|\omega) = f_j^e(s_j^e|\omega) \text{ for all } j \leq t; f^{e'}(s^*, s^*, \ldots, s^*|\omega) = f^e(\mathbf{s}^e|\omega)$$

and $\forall \mathbf{s} \neq \mathbf{s}^*, \mathbf{s} \in \{s^*, s^{-*}\}^t$,

$$f^{e'}(\mathbf{s}|\omega) = \sum_{\substack{\mathbf{s}' \in S^e \text{ s.t.} \\ \text{if } s_j = s^* \text{ then } s'_j = s_j^e \\ \text{otherwise } s'_j \neq s_j^e}} f^e(\mathbf{s}'|\omega)$$

Note that by $(*)$, e' is an explanation of $\mathbf{q}^t$, as the marginal distributions of the observed signals are maintained. Moreover, by $(*)$, it is an explanation which maintains the same likelihood of observing $\mathbf{q}^t$ as e induces.■

Proof of Lemma 2: We start by showing that a forecast $q_h = \beta p + (1-\beta) q_l$ can be ignored if $q_l \in I(\mathbf{q}^t)$, for any $\beta \in (0, 1]$. To see this, note first that $\omega^h = \omega^l$: this arises as $\arg\max_\omega \frac{q_h(\omega)}{p(\omega)} = \arg\max_\omega \frac{\beta p(\omega) + (1-\beta) q_l(\omega)}{p(\omega)} = \arg\max_\omega (\beta + (1-\beta)\frac{q_l(\omega)}{p(\omega)}) = \arg\max_\omega \frac{q_l(\omega)}{p(\omega)}$. Then note that for any ω, $\frac{\alpha_h(\omega)}{\bar{\alpha}_h} = \frac{\frac{q_h(\omega)}{p(\omega)}}{\frac{q_h(\omega^l)}{p(\omega^l)}} = \frac{\frac{\beta p(\omega) + (1-\beta) q_l(\omega)}{p(\omega)}}{\frac{\beta p(\omega^l) + (1-\beta) q_l(\omega^l)}{p(\omega^l)}} = \frac{\beta + (1-\beta)\frac{q_l(\omega)}{p(\omega)}}{\beta + (1-\beta)\frac{q_l(\omega^l)}{p(\omega^l)}} \geq \frac{\frac{q_l(\omega)}{p(\omega)}}{\frac{q_l(\omega^l)}{p(\omega^l)}} = \frac{\alpha_l(\omega)}{\bar{\alpha}_l}$.

Now note that all forecasts in $q_{t+1} \in I(\mathbf{q}^t)$ can be ignored by Proposition 1. As a result of this and the above all forecasts $q_{t+1} = \beta p + (1-\beta) q$ for some $q \in I(\mathbf{q}^t)$ and $\beta \in [0, 1]$ can be ignored.■

Proof of Proposition 2: Suppose that $e_t^* = (\mathcal{I}^{e_t^*}, \mathbf{s}^{e_t^*})$ is a ML explanation of $\mathbf{q}^t$. Recall that by Proposition 1, any ML explanation requires that $f^e(\mathbf{s}^e|\omega) = \min_{j \in \{1,2,..,t\}} \frac{\alpha_j(\omega)}{\bar{\alpha}_j}$.

We now construct a new ML explanation, e_{t+1}^* for $\mathbf{q}^{t+1}$ which is consistent with e_t^*. To construct $e_{t+1}^* = (\mathcal{I}^{e_{t+1}^*}, \mathbf{s}^{e_{t+1}^*})$ we set:

(i) $S^{e_{t+1}^*} = \times_{j=1}^{t+1} S_j^{e_{t+1}^*}$, with $\times_{j=1}^{t} S_j^{e_{t+1}^*} = S^{e_t^*}$, $S_{t+1}^{e_{t+1}^*} = \{s_{t+1}^{e_{t+1}^*}, s_{t+1}^{e_{t+1}^{-*}}\}$ and $\mathbf{s}^{e_{t+1}^*} = (s_1^{e_{t+1}^*}, \ldots s_t^{e_{t+1}^*}, s_{t+1}^{e_{t+1}^*})$ such that $(s_1^{e_{t+1}^*}, \ldots, s_t^{e_{t+1}^*}) = \mathbf{s}^{e_t^*}$.

(ii) $f_j^{e_{t+1}^*}(s_j^{e_{t+1}^*}|\omega) = \frac{\alpha_j(\omega)}{\bar{\alpha}_j}$ for all $j \leq t+1$.

(iii) The marginal of $f^{e_{t+1}^*}(.|\omega)$ on $S^{e_t^*}$ to equal $f^{e_t^*}(.|\omega)$.

We now set two cumulative marginals: on the first t signals, and on the $t+1th$ signal. Order signals in $S_j^{e_{t+1}^*}$ as a normalisation so that $s_j^{e_{t+1}^*}$ is the smallest in $S_j^{e_{t+1}^*}$. This, and (i)-(iii), implies that we have:

$$\begin{aligned} F_{t+1}^{e_{t+1}^*}(s_{t+1}^{e_{t+1}^*}|\omega) &= f_{t+1}^{e_{t+1}^*}(s_{t+1}^{e_{t+1}^*}|\omega), \quad F_{t+1}^{e_{t+1}^*}(s_{t+1}^{e_{t+1}^{-*}}|\omega) = 1 \\ F^{e_{t+1}^*}(\mathbf{s}^{e_t^*}|\omega) &= f^{e_{t+1}^*}(\mathbf{s}^{e_t^*}|\omega) = f^{e_t^*}(\mathbf{s}^{e_t^*}|\omega) \end{aligned}$$

where the remainder of $F^{e^*_{t+1}}(\mathbf{s}^t|\omega)$ on $\mathbf{s}^t \in S^{e^*_t}$, $\mathbf{s}^t \neq \mathbf{s}^{e^*_t}$, can be completed using (iii).

We conclude the construction by setting the joint distribution on $S^{e^*_{t+1}}$, combining the marginals $F^{e^*_{t+1}}(\mathbf{s}^t|\omega)$ and $F^{e^*_{t+1}}_{t+1}(s_{t+1}|\omega)$. We first consider $\mathbf{s}^{e^*_{t+1}}$. For each state ω, and $\mathbf{s}^{e^*_{t+1}}$, we set:

$$\begin{aligned}(*)F^{e^*_{t+1}}(\mathbf{s}^{e^*_{t+1}}|\omega) &= f^{e^*_{t+1}}(\mathbf{s}^{e^*_{t+1}}|\omega) = \\ &= \min\{f^{e^*_{t+1}}_{t+1}(s^{e^*_{t+1}}_{t+1}|\omega), f^{e^*_{t+1}}(\mathbf{s}^{e^*_t}|\omega)\} = \min\{F^{e^*_{t+1}}_{t+1}(s^{e^*_{t+1}}_{t+1}|\omega), F^{e^*_{t+1}}(\mathbf{s}^{e^*_t}|\omega)\}\end{aligned}$$

Next, as we did in Proposition 1, for all other $\mathbf{s} = \{\mathbf{s}^t, s_{t+1}\} \neq \mathbf{s}^{e^*_{t+1}}$, we set

$$F^{e^*_{t+1}}(\mathbf{s}|\omega) = \min\{F^{e^*_{t+1}}_{t+1}(s_{t+1}|\omega), F^{e^*_t}(\mathbf{s}^t|\omega)\}.$$

This will be a proper cdf as there is always a joint information structure which achieves the upper Frechet bound.

We have therefore constructed an explanation e^*_{t+1} which: *(a)* explains the data by (ii); *(b)* maximises the likelihood of the data, as by $(*)$, (ii) and (iii), $f^{e^*_{t+1}}(\mathbf{s}^{e^*_{t+1}}|\omega) = \min_{j\in\{1,2,..,t+1\}} \frac{\alpha_j(\omega)}{\bar{\alpha}_j}$; *(c)* is time-consistent by construction, following (iii).■

Proof of Proposition 3: In text.■

Proof of Proposition 4: (i) This is a corollary of Proposition 1 and the law of large numbers. (ii) In what follows we will denote a forecast about the state by $q \in [0,1]$, interpreted as the probability of state 1, where p is the prior probability that the state is 1. We will construct a true signal generating process by choosing the cumulative distributions over posteriors it generates, $F(q|1)$ in state 1 and $F(q|0)$ in state 0, with corresponding continuous densities $f(q|1)$ and $f(q|0)$.

Let us focus on state $\omega = 0$ and hence on $F(q|0)$. We consider a distribution that has symmetric tails so that $F(q|0) = 1 - F(1-q|0)$ for all $q > \hat{q}$ for some $\hat{q} > 0.5$. To approximate the limit distribution of the extreme posteriors we note that these are the extreme values of $F(q|0)$. We can therefore use the extreme value limit results in Fisher and Tippett (1928) and Leadbetter et al (1983). In particular, we construct $F(q|0)$ to satisfy the following:

$$\exists \gamma(q) > 0 \text{ such that } \lim_{q\to 1} \frac{1 - F(q + x\gamma(q)|0)}{1 - F(q|0)} = e^{-x} \text{ for any } x \in R.$$

By Leadbetter et al (1983) for distributions that satisfy the above condition, there exists sequences $\{a_n\}$, $\{b_n\}$ such that $a_n = \gamma(F^{-1}(1-\frac{1}{n})) \to 0$ and $b_n = F^{-1}(1-\frac{1}{n}) \to 1$ so that we can use the Gumbel distribution to approximate the distribution of the maximal posterior of $F(q|0)$, maxV_n:

$$\frac{\max V_n - b_n}{a_n} \sim \exp\{-\exp(-\frac{\max V_n - b_n}{a_n})\},$$

So if we choose two cutoffs $0 < \alpha_n < \beta_n < 1$ such that $\frac{\beta_n - b_n}{a_n} \to \beta$, $\frac{\alpha_n - b_n}{a_n} \to \alpha$ and $\beta > \alpha$ we have that the probability $\max V_n \in [\alpha_n, \beta_n]$ converges to $\exp\{-\exp(\beta)\} - \exp\{-\exp(\alpha)\} > 0$. Similarly, by symmetry, for the minimum value $\min V_n$ we will have that the probability $\min V_n \in [1-\beta_n, 1-\alpha_n]$ converges to $\exp\{-\exp(\beta)\} - \exp\{-\exp(\alpha)\}$.

Note that as $\frac{1-F(q+x\gamma(q)|0)}{1-F(q|0)} \to_{q\to 1} e^{-x}$ for any $x \in R$ this implies that for large enough q, $q + x\gamma(q) \leq 1$, and so we have $\gamma(q) < \frac{1-q}{x}$ for any $x \in R$. As a result we have that,

$$\frac{1-\beta_n}{1-\alpha_n} \simeq \frac{1-b_n-\beta a_n}{1-b_n-\alpha a_n} = \frac{\frac{1-b_n}{a_n} - \beta}{\frac{1-b_n}{a_n} - \alpha} \to 1$$

as for large enough q, $\frac{1-b_n}{a_n} = \frac{1-F^{-1}(1-\frac{1}{n})}{\gamma(F^{-1}(1-\frac{1}{n}))} > x\frac{1-F^{-1}(1-\frac{1}{n})}{1-F^{-1}(1-\frac{1}{n})} = x$ for any $x \in R$.

Therefore, with probability $(\exp\{-\exp(\beta)\} - \exp\{-\exp(\alpha)\})^2$ we have that

$$1 \leftarrow \frac{1-\beta_n}{1-\alpha_n} < \frac{1-\max V_n}{\min V_n} < \frac{1-\alpha_n}{1-\beta_n} \to 1$$

Moreover, note that $\frac{1-\min V_n}{\max V_n} \to 1$. Therefore, there is a strictly positive probability that the prior will matter for the limit prediction, or in other words, with probability $(\exp\{-\exp(\beta)\} - \exp\{-\exp(\alpha)\})^2$,

$$\mu = \frac{1}{1 + \frac{1-p}{p}\frac{1-\max V_n}{\min V_n}\frac{1-\min V_n}{\max V_n}} \to_{n\to\infty} p.$$

Finally, we need to check that the above can indeed be constructed as distribution over posteriors. A distribution over posteriors arising from Bayesian updating implies a joint restriction on $f(q|1)$ and $f(q|0)$ so that for any q:

$$q = \frac{pf(q|\omega=1)}{pf(q|\omega=1) + (1-p)f(q|\omega=0)} \Rightarrow f(q|1) = f(q|0)\frac{(1-p)}{p}\frac{q}{1-q}.$$[12]

When for example considering the "tails" of $f(q|1)$ and $f(q|0)$, the above means that when $q \to 0$, assuming that $f(q|0) < \infty$ then $f(q|1) \to 0$ at a certain rate. Similarly, $f(q|0) = f(q|1)\frac{p}{1-p}\frac{1-q}{q}$ so that, assuming that $f(q|1) < \infty$, as $q \to 1$ then $f(q|0) \to 0$ at a certain rate. However, given that our approximation implies that $f(q|0)$ drops to zero around the tails very quickly, specifically, as $f(q|0) \to_{q\to 1} 0$ faster than $\frac{q}{1-q} \to_{q\to 1} \infty$, then we can indeed construct $f(q|1)$ as a probability distribution so that they can jointly satisfy the above Bayesian restriction.■

References

[1] Acemoglu, D., V. Chernozhukov and M. Yildiz (2016), Fragility of asymptotic agreement under Bayesian learning, Theoretical Economics 11, 187-225.

[12] These conditions in turn imply Bayesian plausability, i.e., that $\int_0^1 q(pf(q|1) + (1-p)f(q|0))dq = p$.

[2] Arieli, I., J. Babichenko and R. Smorodinsky (2019), Robust Forecast Aggregation, PNAS.

[3] Berger, J., Berliner, L.M. (1986), Robust Bayes and Empirical Bayes Analysis with Contaminated Priors, The Annals of Statistics, 461-486.

[4] Cherry, J. and Y. Salant (2018), Statistical Inference in Games, mimeo.

[5] De Filippis, R., A. Guarino, P. Jehiel and T. Kitagawa (2017), Updating ambiguous beliefs in a social learning experiment, mimeo.

[6] De Marzo, PM, D. Vayanos, J. Zwiebel (2003). Persuasion bias, social influence and unidimensional opinions. Q. J. Econ. 118:909–68.

[7] Ellis, A. and M. Piccione (2017), Correlation Misperception in Choice, American Economic Review 107(4):1264-92.

[8] Enke, B. and F. Zimmerman (2019), Correlation Neglect in Belief Formation, forthcoming, Review of Economics Studies 86(1), 313–332.

[9] Epstein, L. G. and M. Schneider (2007), Learning Under Ambiguity, Review of Economic Studies 74, 1275–1303.

[10] Eyster, E. and G. Weizsäcker (2011). Correlation Neglect in Financial Decision-Making. Discussion Papers of DIW Berlin 1104.

[11] Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. Messenger of Mathematics 41 155-160.

[12] Fisher, R.A. and Tippett, L.H.C. (1928), Limiting forms of the frequency distribution of the largest and smallest member of a sample, Proc. Camb. Phil. Soc., 24 (2): 180–190.

[13] Gilboa, I. and D, Schmeidler (2003), Inductive Inference: An Axiomatic Approach, Econometrica, Vol. 71, No. 1. pp. 1-26.

[14] Gilboa, I. and D, Schmeidler (2010), Simplicity and likelihood: An axiomatic approach, Journal of Economic Theory, Elsevier, vol. 145(5), pages 1757-1775

[15] Glatt, E. (2013), The Unanimous Verdict According to the Talmud: Ancient Law Providing Insight into Modern Legal Theory, Pace International Law Review, Vol 3, No. 10.

[16] Glaeser, E. and C. R. Sunstein (2009), Extremism and social learning, Journal of Legal Analysis, Volume 1, Number 1.

[17] Golub, B. and M. Jackson (2012), How Homophily Affects the Speed of Learning and Best-Response Dynamics, Quarterly Journal of Economics, pp. 1287–1338.

[18] Good, I. (1965), The Estimation of Probabilities: An Essay on Modern Bayesian Methods. The MIT Press.

[19] Gunn, L.J., F.s Chapeau-Blondeau, M.D. McDonnell, Bruce R. Davis, Andrew Allison, Derek Abbott (2016), Too good to be true: when overwhelming evidence fails to convince, Proceedings of the Royal Society A. 472 20150748.

[20] Hossain, T. and R. Okui (2018), Belief Formation Under Signal Correlation, mimeo.

[21] Joe, H., Multivariate Models and Dependence Concepts, Chapman & Hall, London, (1997).

[22] Kallir, I. and Sonsino, D. (2009). The Perception of Correlation in Investment Decisions. Southern Economic Journal 75 (4): 1045-66.

[23] Leadbetter, M. R., Lindgren, G. and Rootze H., Extremes and Related Properties of Random Sequences and Processes, (1983). New York: Springer Verlag.

[24] Levy, G., Moreno de Barreda I. and R. Razin (2018), Persuasion with Correlation Neglect, mimeo, LSE.

[25] Levy, G. and R. Razin (2012), When do Simple Policies Win?. Economic Theory (Political Economy Special Issue), April 2012, 49(3).

[26] Levy, G. and R. Razin (2013), Dynamic Legislative Decision Making when Interest Groups Control the Agenda, Journal of Economic Theory, 2013, vol. 148(5), 1862-1890.

[27] Levy, G. and R. Razin (2015), Correlation Neglect, Voting Behaviour and Information Aggregation, with Ronny Razin, American Economic Review.

[28] Levy, G. and R. Razin (2018a), Information Diffusion in Networks with the Bayesian Peer Influence Heuristic, Games and Economic Behaviour, Volume 109, pp: 262-270.

[29] Levy, G. and R. Razin (2018b), Combining Forecasts in the Presence of Ambiguity over Correlation Structures, mimeo, LSE.

[30] Ortoleva, P. (2012), Modeling the Change of Paradigm: Non-Bayesian Reactions to Unexpected News, American Economic Review 2012, 102(6): 2410–2436.

[31] Ortoleva, P. and E. Snowberg (2015), Overconfidence in political economy, American Economic Review, 105: 504-535.

[32] Osborne, M.J., J.S. Rosenthal, and M.A. Turner (2000), Meetings with costly participation, American Economic Review 90, 927–943.

[33] Rabin, M. and D. Vayanos (2010), The Gambler's and Hot-Hand Fallacies: Theory and Applications Review of Economic Studies 77, 730–778.

[34] Schacter, D., J. Harbluk and D. McLachlan (1984), Retrieval without Recollection: An Experimental Analysis of Source Amnesia, Journal of Verbal Learning and Verbal Behaviour 23, 593-611.

[35] Spiegler, R. (2016), Bayesian Networks and Boundedly Rational Expectations. *Quarterly Journal of Economics*, 131 (3) pp. 1243-1290.

[36] Suleymanov, E. (2018), Robust Maximum Likelihood Updating, working paper, department of economics, University of Michigan.