



HAL
open science

Optimally Imprecise Memory and Biased Forecasts

Rava Azeredo da Silveira, Yeji Sung, Michael Woodford

► **To cite this version:**

Rava Azeredo da Silveira, Yeji Sung, Michael Woodford. Optimally Imprecise Memory and Biased Forecasts. 2021. hal-03033626v1

HAL Id: hal-03033626

<https://hal.science/hal-03033626v1>

Preprint submitted on 10 Nov 2021 (v1), last revised 9 Dec 2020 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Optimally Imprecise Memory and Biased Forecasts

Rava Azeredo da Silveira

Département de Physique
Ecole Normale Supérieure
24 rue Lhomond
75005 Paris, France
rava@ens.fr

Yeji Sung

Department of Economics
Columbia University
420 W. 118th Street
New York, NY 10027
ys2992@columbia.edu

Michael Woodford*

Department of Economics
Columbia University
420 W. 118th Street
New York, NY 10027
mw2230@columbia.edu

*corresponding author

Optimally Imprecise Memory and Biased Forecasts*

Rava Azeredo da Silveira
ENS and University of Basel

Yeji Sung
Columbia University

Michael Woodford
Columbia University

November 12, 2020

Abstract

We propose a model of optimal decision making subject to a memory constraint. The constraint is a limit on the complexity of memory measured using Shannon’s mutual information, as in models of rational inattention; but our theory differs from that of Sims (2003) in not assuming costless memory of past cognitive states. We show that the model implies that both forecasts and actions will exhibit idiosyncratic random variation; that average beliefs will also differ from rational-expectations beliefs, with a bias that fluctuates forever with a variance that does not fall to zero even in the long run; and that more recent news will be given disproportionate weight in forecasts. We solve the model under a variety of assumptions about the degree of persistence of the variable to be forecasted and the horizon over which it must be forecasted, and examine how the nature of forecast biases depends on these parameters. The model provides a simple explanation for a number of features of reported expectations in laboratory and field settings, notably the evidence of over-reaction in elicited forecasts documented by Afrouzi *et al.* (2020) and Bordalo *et al.* (2020a).

*We thank Hassan Afrouzi, Ben Hébert, David Laibson, Yueran Ma, and Andrei Shleifer for helpful discussions, and are especially grateful to the authors of Afrouzi *et al.* (2020) for sharing the data plotted in Figure 7. We also thank the Alfred P. Sloan Foundation, the CNRS through UMR 8023, and the IOB for research support.

The hypothesis of rational expectations (RE) proposes that decisions are based on expectations that make use of all available information in an optimal way: that is, those that would be derived by correct Bayesian inference from an objectively correct prior and the data that has been observed to that date. Yet both in surveys of individual forecasts of macroeconomic and financial variables and in forecasts elicited in experimental settings, beliefs are more heterogeneous than this hypothesis should allow, and forecast errors are predictable on the basis of variables observable by the forecasters, contrary to this hypothesis. In particular, a number of studies have argued that forecasts typically over-react to new realizations of the variable being forecasted. (See Bordalo *et al.*, 2020a, and Afrouzi *et al.*, 2020, for recent examples with extensive references to prior literature.)

A variety of models of expectation formation have been proposed that allow for such over-reaction. The simplest type of model simply posits that the forecasted future value of a variable is a particular linear function of the last few observations of the variable; with an appropriate choice of the coefficients (such as those proposed by Metzler, 1941), a forecasting heuristic of this kind may imply that a recent increase in the variable will be extrapolated into the future, so that further increases are anticipated, regardless of whether the degree of serial correlation of changes in the variable make this an optimal forecast. A classic critique of such proposals, however, is that of Muth (1961): why should decision makers continue to forecast in this way, if their forecasts are systematically biased, as repeated observations should eventually make clear? Moreover, a mechanical heuristic with fixed coefficients is unable to explain how the biases in observed forecasts change depending on the persistence of the process that is forecasted (Afrouzi *et al.*, 2020).

Fuster *et al.* (2010, 2011) propose a more sophisticated model, in which decision makers are assumed to forecast a time series by modeling it as a low-order autoregressive process; the coefficients of their forecasting rule are those implied by the AR(k) model that best fits the autocorrelation function of the actual series, for some fixed bound on k . The authors argue that actual time series often involve long-horizon dependencies, and show that in this case (say, an AR(40) process forecasted by people who consider models with no more than 10 lags), long-horizon forecasts using the best-fitting AR(k) model can significantly over-react to recent trends in the data.

This proposal, however, remains subject to several objections. Why should the restriction to models of the data with a fixed upper bound on k be maintained, even when the available sequence of observations with which to estimate the model becomes unboundedly long? Moreover, even if one grants that a constraint on model complexity requires that no more than some finite number of explanatory variables be stored and used as a basis for forecasts, why must the explanatory variables correspond to the last k observations of the series? In the kind of example in which Fuster *et al.* argue that their proposal predicts over-reaction, more accurate long-horizon forecasts would be possible if the forecast were conditioned on a long moving average of observations, rather than only recent observations; yet tracking a small number of moving averages would seem no more complex than always having access to the last k observations. And above all, the Fuster *et al.* explanation implies that over-reaction should only be observed in the case of variables that are not well-described by an AR(k) process of low enough order. Yet Afrouzi *et al.* (2020) find significant over-reaction in an experiment in which the true data-generating process is an AR(1) process; and in fact, they find the most severe degree of over-reaction (as discussed further below) when the process

to be forecasted is white noise.

Here we offer a different explanation for the pervasiveness of over-reaction. We consider a model in which a decision maker’s forecasts (or more generally, actions with consequences that depend on the future realization of some variable) can be based both on currently observable information and an imperfect memory of past observations. Subject to this constraint on the information that the decision rule can use, we assume that the rule is optimal. Moreover, rather than making an arbitrary assumption about the kind of statistics about past experience that can be recalled with greater or lesser precision, we allow the memory structure to be specified in a very flexible way, and assume that it is optimized for the particular decision problem, subject only to a constraint on the overall complexity of the information that can be stored in (and retrieved from) memory — or more generally, subject to a cost of using a more complex memory structure.

In the limiting case in which the cost of memory complexity is assumed to be negligible, the predictions of our model coincide with those of the rational expectations hypothesis. But when the cost is larger (or the constraint on memory complexity is tighter), our model predicts that forecasts should be both heterogeneous (even in the case of forecasters who observe identical data) and systematically biased. Moreover, the predicted biases include the type of over-reaction to news documented in surveys of forecasts of macroeconomic and financial time series by Bordalo *et al.* (2020a) and in laboratory forecasting experiments by Afrouzi *et al.* (2020). And unlike the model of Fuster *et al.* (2010, 2011), our model predicts that over-reaction to news will be most severe in the case of time series exhibiting little serial correlation.

In seeking to endogenize the information content of the noisy cognitive state on the basis of which people must act, our theory is in the spirit of Sims’s (2003) theory of “rational inattention”; and indeed, we follow Sims in modeling the complexity constraint using information theory. There is nonetheless an important difference between our theory and that of Sims (2003). Sims assumes a constraint on the precision with which new observations of the world can reflect any current or past conditions outside the head of the decision maker, but assumes perfectly precise memory of all of the decision maker’s own past cognitive states, and also assumes that past external states can be observed at any time with the same precision as current conditions. We instead assume (for the sake of simplicity) that the current external state can be observed with perfect precision, but that memory of past cognitive states is subject to an information constraint; and we further assume that the decision maker has no access to external states that occurred in the past, except through (information-constrained) access to her own memory of those past states. These differences are crucial for the ability of our model to explain over-reaction to news.

Other recent papers that explore the consequences of assuming that memory allows only a noisy recollection of past observations include Neligh (2019) and Afrouzi *et al.* (2020). While these authors also assume that some aspects of memory structure are optimized for a particular decision problem, the classes of memory structures that they consider are different than the one that we analyze here.

Both of these papers assume that successive observations of the external state are individually encoded (possibly with variable precision) and stored in memory at the time of the observation; the precision of the record of each observation then evolves over time in a way that is exogenously specified, rather than optimized; and finally, when memory is accessed

later to make a decision, the nature of the signal about the contents of memory may also be endogenized. (Neligh focuses on endogenizing the precision of the initial encoding of individual observations; Afrouzi *et al.* instead focus on endogenizing the precision of what is retrieved from memory.) Both papers take it as given that memory is a vector of separately encoded values of individual observations, and allow no re-encoding of the contents of memory between the time of an initial observation and its retrieval for use in a decision. Our concern is instead with the way in which it is optimal for memory to be constantly re-encoded as time passes, in order to make the most efficient use of a finite limit on the complexity of the stored representations. We comment further on the differences between our framework and those of these other papers below.¹

We present the assumptions of our model and state the optimization problem that we consider in section 1. Section 2 offers a general characterization of the optimal memory structure in our model, showing in particular that it is optimal under our assumptions for the memory state at each point in time to be represented by a single real number, a random variable the mean of which depends on the entire sequence of previous observations. Section 3 illustrates the model’s implications, discussing quantitative aspects of numerical solutions of the model for particular parameter values. We emphasize the failure of beliefs ever to converge to those associated with a rational expectations equilibrium, and show that instead, there are perpetual stationary fluctuations in subjective beliefs similar (though not identical) to those predicted by models of “constant-gain learning” (Evans and Honkapohja, 2001). Finally, section 4 presents the quantitative predictions of the model for statistics of the kind reported by Afrouzi *et al.* (2020) and Bordalo *et al.* (2020a), showing not only that the model can produce over-reaction to news, but that it can be parameterized so as to predict roughly the degree of over-reaction measured by these authors. Section 5 concludes.

1 A Flexible Model of Imprecise Memory

Here we introduce the class of linear-quadratic-Gaussian decision problems that we study, and specify the nature of a general constraint on the precision of memory. This gives rise to a dynamic optimization problem, the solution to which we study below.

1.1 The forecasting problem

In the kind of decision problem which we consider, a decision maker [DM] observes the successive realizations of a univariate stochastic process y_t (“the external state”), which we assume to be a stationary AR(1) process. We write the law of motion of this process in the form

$$y_t = \mu + \rho(y_{t-1} - \mu) + \epsilon_{yt}, \quad (1.1)$$

where μ is the mean value of the external state, ρ is the coefficient of serial correlation (with $|\rho| < 1$), and $\{\epsilon_{yt}\}$ is an i.i.d. sequence, drawn each period from a Gaussian distribution

¹See section 1.2. In addition to considering a different class of possible memory structures, Neligh (2019) addresses largely distinct questions from those analyzed here. Afrouzi *et al.* (2020) instead adopt the explanation that we propose here for their (previously circulated) experimental findings, but show that similar biases are also predicted by a simpler model of noisy memory than the one that we present here.

$N(0, \sigma_\epsilon^2)$. The variance of the external state (conditional on the value of μ and the other parameters) will therefore equal $\sigma_y^2 \equiv \sigma_\epsilon^2 / (1 - \rho^2)$.

The DM's problem is to produce each period a vector of forecasts z_t , so as to minimize the expected value of a discounted quadratic loss function

$$\mathbb{E} \sum_{t=0}^{\infty} \beta^t (z_t - \tilde{z}_t)' W (z_t - \tilde{z}_t), \quad (1.2)$$

where $0 < \beta < 1$, W is a positive definite matrix specifying the relative importance of accuracy of the different dimensions of the vector of forecasts, and the eventual outcomes that the DM seeks to forecast are functions of the future evolution of the external state,²

$$\tilde{z}_t \equiv \sum_{j=0}^{\infty} A_j y_{t+j},$$

where the coefficients $\{A_j\}$ satisfy $\sum_j |A_j| < \infty$. This formalism allows us to assume that the DM may produce forecasts about the future state at multiple horizons (as is typically true in surveys of forecasters, and also in the experiment of Afrouzi *et al.*, 2020). It also allows us to treat cases in which the DM may choose a vector of actions, the rewards from which are a quadratic function of the action vector and the external state in various periods; the problem of action choice to maximize expected reward in such a case is equivalent to a problem of minimizing a quadratic function of the DM's error in forecasting certain linear combinations of the value of the external state at various horizons.³

To simplify our discussion, we assume that the second moments of the stochastic process for the external state are known (more precisely, that the DM's decision rule can be optimized for particular values of these parameters, that are assumed to be the correct ones), while the first moment is not, so that the DM's decision rule must respond adaptively to evidence about the unknown mean value provided by the DM's observations of the state. Thus the values of the parameters ρ and σ_ϵ^2 are assumed to be known, while μ is not; the parameter μ is assumed to be drawn from a non-degenerate prior distribution, $\mu \sim N(0, \Omega)$. Conditional on the value of μ , the initial lagged state y_{-1} is assumed to be drawn from the prior distribution $N(\mu, \sigma_y^2)$, the ergodic distribution for the external state given a value for μ . When we consider the optimality of a possible decision rule for the DM, we integrate over this prior distribution of possible values for μ and y_{-1} , assuming that the decision rule must operate in the same way regardless of which values happen to be true in a given environment.

Note that in the absence of any memory limitation — and given the assumption (maintained in this paper) of perfect observability of the realizations of y_t — it should be possible eventually for the DM to learn the value of μ to arbitrary precision, so that despite our

²Note that the variables denoted \tilde{z}_t are not quantities the value of which is determined at time t ; the subscript t is used to identify the time at which the DM must produce a forecast of the quantity, not the time at which the outcome will be realized. Thus the best possible forecast of \tilde{z}_t at time t , even with full information, would be given by $\mathbb{E}_t \tilde{z}_t$, which will generally not be the same as the realized value \tilde{z}_t .

³For example, in a standard consumption-smoothing problem with quadratic consumption utility, the DM's level of expected utility depends on the accuracy with which "permanent income" is estimated at each point in time. This requires the DM to forecast a single variable \tilde{z}_t , for which the coefficient A_j is proportional to β^j for all $j \geq 0$.

assumption that the value of μ must be learned, the optimal decision rule should coincide asymptotically with the full-information rational-expectations prediction. We show, however, that this is not true when the precision of memory is bounded.

In any problem of this form (regardless of the assumed memory limitations, which have yet to be specified), the DM's problem can equivalently be formulated as one of simply choosing an estimate $\hat{\mu}_t$ of the unknown mean μ at each date t , based on the information available at the time that z_t must be chosen. It follows from the law of motion (1.1) that

$$E_t \tilde{z}_t = \sum_{j=0}^{\infty} A_j [\mu + \rho^j (y_t - \mu)],$$

where we use the notation $E_t[\cdot]$ for the expected value conditional on the true state at time t , i.e., the value of μ and the history of realizations (y_0, \dots, y_t) , even though not all of this information is available to the DM. Conditioning instead on the coarser information set that represents the DM's cognitive state at time t (and noting that this includes precise awareness of the value of y_t), we similarly find that the optimal estimate of \tilde{z}_t will be given by

$$z_t = \sum_{j=0}^{\infty} A_j [\hat{\mu}_t + \rho^j (y_t - \hat{\mu}_t)], \quad (1.3)$$

where $\hat{\mu}_t$ is the expectation of μ conditional on the DM's information set at time t .

We show in the appendix that the DM's expected loss cannot be reduced by restricting attention to a class of decision rules of the form (1.3), under different possible assumptions about how the estimate $\hat{\mu}_t$ is formed.⁴ In the case of any forecasting rule of that kind, the loss function (1.2) is equal to

$$\alpha \cdot \sum_{t=0}^{\infty} \beta^t MSE_t \quad (1.4)$$

plus a term that is independent of the DM's forecasts, where

$$MSE_t \equiv E[(\hat{\mu}_t - \mu)^2]$$

is the mean squared error in estimating μ , and $\alpha > 0$ is a constant that depends on the coefficients $\{A_j\}$ and W . Thus one can equivalently formulate the DM's problem as one of optimal choice of an estimate $\hat{\mu}_t$ each period, so as to minimize MSE_t .

1.2 Memory structures and their cost

We assume that the memory carried into each period $t \geq 0$ can be summarized by a vector m_t of dimension d_t ; the action chosen in period t (i.e., the choice of $\hat{\mu}_t$) must be a function of the cognitive state specified by $s_t = (m_t, y_t)$. The dimension of the memory state is assumed only to be finite, and is not required to be the same for all t . (The case of *perfect memory* can be accommodated by our notation, by assuming that $d_t = t$, and that the elements of the vector m_t correspond to the values $(y_0, y_1, \dots, y_{t-1})$.) We assume that current external

⁴See Appendix A for details of the argument.

state y_t is perfectly observable,⁵ but that behavior can depend on past states only to the extent that memory provides information about them.

We further suppose that the memory state evolves according to a linear law of motion of the form

$$m_{t+1} = \Lambda_t s_t + \omega_{t+1}, \quad \omega_{t+1} \sim N(0, \Sigma_{\omega,t+1}) \quad (1.5)$$

starting from an initial condition of dimension $d_0 = 0$ (that is, s_0 consists only of y_0). However, the dimension d_{t+1} of the memory that is stored, and the elements of the matrices $\Lambda_t, \Sigma_{\omega,t+1}$ are allowed to be arbitrary; we require only that $\Sigma_{\omega,t+1}$ must be positive semi-definite (though it need not be of full rank).

For example, one type of memory structure that this formalism allows us to consider is an “episodic” memory of the kind assumed by Neligh (2019).⁶ In this case, $d_t = t$, and there is an element of m_t corresponding to each of the past observations y_τ for $0 \leq \tau \leq t - 1$ (generalizing the case of perfect memory just discussed). The memory of y_τ at some later time t is given by $m_{\tau+1,t} = y_\tau + u_{\tau+1,t}$, where $u_{\tau+1,t}$ is a Gaussian noise term, independent of the value of y_τ , and with a variance that is necessarily non-decreasing in t . This can be represented by letting $d_t = t$, Λ_t be the identity matrix of dimension $t + 1$, and $\Sigma_{\omega,t+1}$ a diagonal matrix of dimension $n + 1$ (with non-negative elements, but not necessarily of full rank).

Another type of memory that we can consider is one in which only the n most recent past observations of the external state can be recalled, though these are recalled with perfect precision. The requirement that forecasts be functions of the cognitive state would then require them to be functions of $(y_t, y_{t-1}, \dots, y_{t-n})$ for some finite n , as under the hypothesis of “natural expectations” proposed by Fuster, Hébert, and Laibson (2011). This case would correspond to a specification in which $d_t = n$ for all t ; Λ_t is an $n \times (n + 1)$ matrix, the right $n \times n$ block of which is an identity matrix, and the first column of which consists entirely of zeroes; and $\Sigma_{\omega,t+1} = 0$. Our formalism is much more flexible than either of these cases, however, and neither of those specifications turns out to be optimal.

We limit the precision of memory by further assuming that there is a cost of storing and/or accessing the memory state m_{t+1} , that is an increasing function of the Shannon mutual information between the memory state m_{t+1} and the cognitive state s_t about which it provides information.⁷ If I_t is the mutual information between these two random variables, we assume a memory cost in period t of $c(I_t)$, where $c(I)$ is an increasing, convex function. Two extreme possibilities, both of which we consider further below, are the one in which $c(I)$ is a linear function, $c(I) = \theta \cdot I$ for some $\theta > 0$; and the opposite limiting case in which

⁵We might also assume that the current state is observable only imprecisely, as in the model of Sims (2003); but in the present treatment, we simplify the analysis, and highlight the consequences of imperfect memory, by considering the limiting case in which there is no cost of precise observation of the current external state.

⁶Note however that Neligh’s model is not a special case of ours, because in addition to restricting attention to a more special class of memory structures, he assumes a different cost function for precision than the one we propose below.

⁷Mutual information is a non-negative scalar quantity that can be defined for any joint distribution for (s_t, m_{t+1}) , that measures the degree to which the realized value of either random variable provides information about the value of the other (Cover and Thomas, 2006). This measure is used to determine the relative cost of different information structures in the rational inattention theory of Sims (2003); properties of this measure as an information cost function are discussed in Caplin, Dean and Leahy (2019).

there is a finite upper bound \bar{I} on feasible information transmission, with zero cost for any $I < \bar{I}$. Either of these cases allows us to consider a one-parameter family of possible cost functions, ranging from an arbitrarily loose information constraint (θ near zero, \bar{I} very large) to a prohibitively tight one (θ extremely high, \bar{I} near zero).

The cost $c(I_t)$ can equivalently be viewed as either a cost of storing a memory record with information content I_t (that is then available with perfect precision in period $t + 1$), or a cost of retrieving a signal from memory with information content I_t in period $t + 1$ (while the memory stored in period t is taken to have been a perfect record of the period t cognitive state). These two formulations are identical, given that we assume that only the signal m_{t+1} that is retrieved in period $t + 1$ can be stored for future use; thus only the fidelity with which the retrieved memory m_{t+1} reproduces the cognitive state s_t matters. Under the retrieval-cost interpretation, however, our model remains importantly different from the one proposed by Afrouzi *et al.* (2020), in which memory contains a perfect record of all past observations, but there is a cost of retrieving a precise signal about the contents of memory for use in a decision. That model assumes that past observations can be stored indefinitely with perfect precision, with a limit on the precision of recall becoming relevant only when memory must be consulted; this means that it does not predict “recency bias” as ours does.⁸

The memory structure each period, together with the rule for choosing an estimate $\hat{\mu}_t$ as a function of each period’s cognitive state, are then assumed to be chosen so as to minimize total discounted costs

$$\sum_{t=0}^{\infty} \beta^t [\alpha \cdot MSE_t + c(I_t),] \tag{1.6}$$

taking into account both the cost of less accurate forecasts (1.4) and the cost of greater memory precision. Note that no expectation is needed in this objective, since both MSE_t and I_t are functions of the entire joint probability distribution of possible values for $\mu, m_t, y_t, \hat{\mu}_t$ and m_{t+1} .

2 The Optimal Memory Structure

We turn now to a general characterization of the solution to the dynamic optimization problem just posed.

2.1 Implications of linear-Gaussian dynamics

For any memory structure in the class specified in the previous section, the posterior distribution over possible values of $(\mu, y_{-1}, y_0, \dots, y_{t-1})$ implied by memory state m_t will be a multivariate Gaussian distribution. It is thus fully characterized by specifying a finite set of first and second moments of the posterior associated with the memory state. Moreover, the particular memory state m_t at a given date t can be identified by the associated vector of first moments. For the second moments of the posterior are the same for all possible memory states at any time t : they depend on the matrices $\{\Lambda_\tau, \Sigma_{\omega, \tau+1}\}$ for $\tau < t$, but not on the

⁸See the discussion in section 3.4 below.

history of the external state, or on the history of realizations of the memory noise $\{\omega_{t+1}\}$. In what follows, we therefore use the notation m_t for the vector of posterior means.

Among the state variables about which the memory state may convey information, we are particularly interested in the vector of variables $x_t = (\mu, y_{t-1})'$, which are the states determined prior to period t that are relevant for predicting the external state in periods $\tau \geq t$. Let $\bar{m}_t \equiv E[x_t | m_t]$ be the two elements of the memory state that identify the posterior mean of x_t , and let Σ_t be the 2×2 block of second moments of x_t under this same posterior, so that

$$x_t | m_t \sim N(\bar{m}_t, \Sigma_t).$$

And let us furthermore introduce the vectors

$$e'_1 \equiv [1 \ 0], \quad c' \equiv [1 - \rho \ \rho]$$

to select particular elements of this reduced state vector. Then $e'_1 \bar{m}_t$ is the posterior mean and $e'_1 \Sigma_t e_1$ the posterior variance for μ ; while $c' \bar{m}_t$ is the posterior mean and $c' \Sigma_t c$ the posterior variance for the full-information forecast $E_{t-1} y_t$.

The posterior mean and variance for μ after also observing y_t will then be given by the usual Kalman filter formulas,

$$\hat{\mu}_t \equiv E[\mu | s_t] = e'_1 \bar{m}_t + \gamma_{1t} [y_t - c' \bar{m}_t], \quad (2.1)$$

$$\hat{\sigma}_t^2 \equiv \text{var}[\mu | s_t] = e'_1 \Sigma_t e_1 - \gamma_{1t}^2 [c' \Sigma_t c + \sigma_\epsilon^2], \quad (2.2)$$

where the Kalman gain is equal to⁹

$$\gamma_{1t} = \frac{e'_1 \Sigma_t c}{c' \Sigma_t c + \sigma_\epsilon^2}. \quad (2.3)$$

Since y_t is observed precisely, these formulas completely characterize posterior beliefs in cognitive state s_t about the states x_{t+1} that are relevant for forecasting y_τ for all $\tau > t$. Note that $\hat{\sigma}_t^2$ is necessarily positive (complete certainty about the value of μ cannot be achieved in finite time, even in the case of perfect memory), and must satisfy the upper bound

$$\hat{\sigma}_t^2 \leq \hat{\sigma}_0^2 \equiv \frac{\Omega \sigma_y^2}{\Omega + \sigma_y^2}, \quad (2.4)$$

which corresponds to the degree of uncertainty about μ after observing the external state in the case of no informative memory whatsoever (the DM's situation in period $t = 0$).

Then the average losses from inaccurate forecasting in period t are given by

$$MSE_t = \hat{\sigma}_t^2. \quad (2.5)$$

This determines the value of one of the terms in (1.6) as a function of the posterior uncertainty associated with the memory state each period. We note that the optimal estimate $\hat{\mu}_t$ depends only on \bar{m}_t (not other components of the memory state), and that the average loss implied by this estimate depends only on the posterior uncertainty Σ_t associated with those same two components.

⁹We use a 1 subscript in the notation for this variable because it is the first element of a vector of Kalman gains, defined in the more general formula given in Appendix B.

2.2 The sufficiency of memory of a reduced cognitive state

We further show in the appendix¹⁰ that an optimal memory structure makes the memory state m_{t+1} a function only of the “reduced cognitive state”

$$\bar{s}_t \equiv \begin{bmatrix} \hat{\mu}_t \\ y_t \end{bmatrix} = E[x_{t+1} | s_t]. \quad (2.6)$$

We first note (using (2.1) and the fact that y_t is part of the cognitive state) that the elements of \bar{s}_t are a linear function of s_t . Thus we can choose a representation of the vector s_t in which its elements are made up of two parts, \bar{s}_t and \underline{s}_t , where the elements of \underline{s}_t are uncorrelated with those of \bar{s}_t . We then observe that

$$\bar{m}_{t+1} = E[\bar{s}_t | m_{t+1}].$$

Hence the only aspect of the memory state that matters for \bar{m}_{t+1} , and hence for determining both the optimal estimate $\hat{\mu}_{t+1}$ and the reduced cognitive state \bar{s}_{t+1} , will be the information that m_{t+1} contains about \bar{s}_t .

To the extent that m_{t+1} conveys any information about the elements of \underline{s}_t , this information has no consequences for the DM’s estimates $\hat{\mu}_\tau$ in any periods $\tau \geq t+1$, but it increases the mutual information between s_t and m_{t+1} , and hence the information cost $c(I_t)$. Hence under an optimal information structure, the reduced memory state \bar{m}_t must evolve according to a law of motion of the form

$$\bar{m}_{t+1} = \bar{\Lambda}_t \bar{s}_t + \bar{\omega}_{t+1}, \quad (2.7)$$

where $\bar{\omega}_{t+1} \sim N(0, \Sigma_{\bar{\omega}, t+1})$ is distributed independently of the cognitive state. And in addition, the complete memory state must convey no more information about s_t than what is conveyed by the reduced memory state, so that we can without loss of generality assume that m_{t+1} consists solely of \bar{m}_{t+1} (so that $d_{t+1} = 2$ for all $t \geq 0$).

Finally, the 2×2 matrices $\bar{\Lambda}_t$ and $\Sigma_{\bar{\omega}, t+1}$ must satisfy additional restrictions in order for the reduced memory state defined in (2.7) to be consistent with the normalization

$$E[\bar{s}_t | \bar{m}_{t+1}] = \bar{m}_{t+1}. \quad (2.8)$$

We show in the appendix that this relationship will hold if and only if¹¹

$$\Sigma_{\bar{\omega}, t+1} = (I - \bar{\Lambda}_t) X_t \bar{\Lambda}_t', \quad (2.9)$$

where $X_t \equiv \text{var}[\bar{s}_t]$. Note that (2.6) implies that

$$\text{var}[x_{t+1}] = \text{var}[\bar{s}_t] + \text{var}[x_{t+1} | s_t],$$

from which we see that

$$X_t = X(\hat{\sigma}_t^2) \equiv \begin{bmatrix} \Omega - \hat{\sigma}_t^2 & \Omega \\ \Omega & \Omega + \sigma_y^2 \end{bmatrix}. \quad (2.10)$$

¹⁰See Appendix C for details of the argument.

¹¹See the introductory section of Appendix D for details of the argument.

Thus the matrix X_t depends only on the value of $\hat{\sigma}_t^2$. In addition, (2.4) implies that X_t will be positive semi-definite (p.s.d.), and non-singular (hence positive definite) except in the case that $\hat{\sigma}_t^2 = \hat{\sigma}_0^2$ (the case of a totally uninformative memory state m_t).

In order for it to be possible for (2.9) to hold, the matrix $\bar{\Lambda}_t$ must satisfy certain properties: (a) the matrix $\bar{\Lambda}_t X_t = X_t \bar{\Lambda}_t'$ must be symmetric (so that the right-hand side of (2.9) is also symmetric); and (b) the right-hand side of (2.9) must be a p.s.d. matrix. For any symmetric, positive definite 2×2 matrix X_t , we let $\mathcal{L}(X_t)$ be the set of matrices $\bar{\Lambda}_t$ with these properties. Then in addition to assuming that $\bar{\Lambda}_t \in \mathcal{L}(X_t)$, the variance matrix $\Sigma_{\bar{\omega}, t+1}$ must be given by (2.9).

In the special case in which m_t is completely uninformative, $\hat{\mu}_t$ is proportional to the observation y_t , so that there exists a vector $w \gg 0$ such that $\bar{s}_t = w \cdot y_t$. In this case,

$$X_t = X_0 \equiv [\Omega + \sigma_y^2] w w',$$

and we can show that the requirements stated above are satisfied by a matrix $\bar{\Lambda}_t$ if and only if $\bar{\Lambda}_t w = \lambda_t w$ (w is a right eigenvector), with an eigenvalue satisfying $0 \leq \lambda_t \leq 1$. Since the two elements of \bar{s}_t are perfectly collinear in this case, the only part of the matrix $\bar{\Lambda}_t$ that matters for the evolution of the memory state is the implied vector $\bar{\Lambda}_t w$ (which must be a multiple of w). Thus we can without loss of generality impose the further restriction that if $\hat{\sigma}_t^2 = \hat{\sigma}_0^2$, we will describe the dynamics of the memory state using a matrix $\bar{\Lambda}_t$ of the form

$$\bar{\Lambda}_t = \lambda_t \frac{w w'}{w' w}, \quad (2.11)$$

for some $0 \leq \lambda_t \leq 1$. We now adopt this more restrictive definition of the set $\mathcal{L}(X_0)$ in this special case.¹²

We have now shown that the memory structure for period t is completely determined by a specification of a matrix $\bar{\Lambda}_t \in \mathcal{L}(X_t)$. In any period t , the value of $\hat{\sigma}_t^2$ and hence the matrix X_t will be implied by the choice of memory structure for the periods prior to t . Given a choice of $\bar{\Lambda}_t$, the variance-covariance matrix $\Sigma_{\bar{\omega}, t+1}$ is uniquely determined by (2.9). As shown in the appendix,¹³ this then uniquely determines Σ_{t+1} , indicating the degree of uncertainty implied by the memory state m_{t+1} , which then determines $\hat{\sigma}_{t+1}^2$ using (2.2). The degree of uncertainty about μ in the following period is then given by a function of the form

$$\hat{\sigma}_{t+1}^2 = f(\hat{\sigma}_t^2, \bar{\Lambda}_t),$$

that is uniquely defined for any non-negative $\hat{\sigma}_t^2$ satisfying the bound (2.4) and any $\bar{\Lambda}_t \in \mathcal{L}(X(\hat{\sigma}_t^2))$.

Then given that we start from an initial degree of uncertainty $\hat{\sigma}_0^2$ at time $t = 0$ defined by (2.4), we can define the class of sequences $\{\bar{\Lambda}_t\}$ for all $t \geq 0$ with the property that $\bar{\Lambda}_t \in \mathcal{L}(X_t)$ for all $t \geq 0$; let us call this class \mathcal{L}^{seq} . Moreover, for any sequence of transition matrices in \mathcal{L}^{seq} , we can uniquely define the sequences of values $\{\Sigma_t, \gamma_{1t}, \hat{\sigma}_t^2, X_t\}$ for all $t \geq 0$

¹²Restricting the set of transition matrices $\bar{\Lambda}_t$ that may be chosen in this way has no consequences for the evolution of the memory state, but it makes equation (2.12) below also valid in the case that $X_t = X_0$, and thus it allows us to state certain conditions below more compactly.

¹³See Appendix D.1 for details of the argument.

implied by it. Thus given any sequence $\{\bar{\Lambda}_t\} \in \mathcal{L}^{seq}$, we can calculate the implied sequence of losses $\{MSE_t\}$ from forecast inaccuracy, using (2.5).

We can also uniquely identify the information cost implied by such a sequence of transition matrices, since as shown in the appendix,¹⁴ the mutual information between s_t and m_{t+1} will be given by

$$I_t = I(\bar{\Lambda}_t) \equiv -\frac{1}{2} \log \det(I - \bar{\Lambda}_t) \quad (2.12)$$

each period. Note that the requirement that $\bar{\Lambda}_t \in \mathcal{L}(X_t)$ implies that

$$0 < \det(I - \bar{\Lambda}_t) \leq 1,$$

so that the quantity (2.12) is well-defined, and necessarily non-negative. As the elements of $\bar{\Lambda}_t$ are made small, so that memory ceases to be very informative about the prior cognitive state, $I - \bar{\Lambda}_t$ approaches the identity matrix, and I_t approaches zero. If $\bar{\Lambda}_t$ is varied in such a way as to make one of its eigenvalues approach 1, $I - \bar{\Lambda}_t$ approaches a singular matrix, and $\Sigma_{\hat{\omega}, t+1}$ must approach a singular matrix as well; this means that in the limit, some linear combination of the elements of \bar{s}_t is a random variable with positive variance that comes to be recalled with perfect precision. In this case, $\det(I - \bar{\Lambda}_t)$ approaches zero, so that I_t grows without bound.

Thus a given sequence of transition matrices $\{\bar{\Lambda}_t\}$ uniquely determines sequences $\{MSE_t, I_t\}$, allowing the value of the objective (1.6) to be calculated. The problem of optimal design of a memory structure can then be reduced to the choice of a sequence $\{\bar{\Lambda}_t\} \in \mathcal{L}^{seq}$ so as to minimize (1.6). This objective is necessarily well-defined for any such sequence, since each of the terms is non-negative; the infinite sum will either converge to a finite value, or will diverge, in which case the sequence in question cannot be optimal.¹⁵

2.3 A recursive formulation

We now observe that if for any point in time t , we know the value of $\hat{\sigma}_t^2$ (which depends on the choices made regarding memory structure in periods $\tau < t$), the set of admissible transition matrices $\{\bar{\Lambda}_\tau\}$ for $\tau \geq t$ specifying the memory structure from that time onward will depend only on the value of $\hat{\sigma}_t^2$, and not any other aspect of choices made about the earlier periods. Moreover, any admissible continuation sequence $\{\bar{\Lambda}_\tau\}$ for $\tau \geq t$ implies unique continuation sequences $\{MSE_\tau, I_\tau\}$ for $\tau \geq t$, so that the value of the continuation objective

$$\sum_{\tau=t}^{\infty} \beta^{\tau-t} [\alpha \cdot MSE_\tau + c(I_\tau)] \quad (2.13)$$

will be well-defined.¹⁶

¹⁴See Appendix D.2 for details of the argument.

¹⁵Note that it is clearly possible to choose memory structures for which the infinite sum converges. For example, if one chooses $\bar{\Lambda}_t = 0$ for all $t \geq 0$ (perfectly uninformative memory at all times), $MSE_t = \hat{\sigma}_0^2$ and $I_t = 0$ each period, and (1.6) will equal the finite quantity $\hat{\sigma}_0^2/(1 - \beta)$.

¹⁶Since a finite value for the continuation objective is always possible (see (2.14) below), it is clear that plans that make (2.13) a divergent series cannot be optimal, and can be excluded from consideration.

We can then consider the problem of choosing an admissible continuation plan $\{\bar{\Lambda}_\tau\}$ for $\tau \geq t$ so as to minimize (2.13), given an initial condition for $\hat{\sigma}_t^2$. (This is simply a more general form of our original problem choosing memory structures for all $t \geq 0$ to minimize (1.6), given the initial condition for $\hat{\sigma}_0^2$ specified in (2.4).) Let $V(\hat{\sigma}_t^2)$ be the lowest achievable value for (2.13), as a function of the initial condition $\hat{\sigma}_t^2$; this function is defined for any value of $\hat{\sigma}_t^2$ satisfying the bound (2.4), and is independent of the date t from which we consider the continuation problem. Note that the lower bound necessarily lies in the interval

$$\alpha\hat{\sigma}_t^2 \leq V(\hat{\sigma}_t^2) \leq \alpha \left[\hat{\sigma}_t^2 + \frac{\beta}{1-\beta} \hat{\sigma}_0^2 \right]. \quad (2.14)$$

(The lower bound follows from the fact that $MSE_t = \hat{\sigma}_t^2$, and all other terms in (2.13) must be non-negative; the upper bound is the value of (2.13) if one chooses $\bar{\Lambda}_\tau = 0$ for all $\tau \geq t$, which is among the admissible continuation plans.)

This value function also necessarily satisfies a Bellman equation of the form

$$V(\hat{\sigma}_t^2) = \min_{\bar{\Lambda}_t \in \mathcal{L}(X(\hat{\sigma}_t^2))} [\alpha\hat{\sigma}_t^2 + c(I(\bar{\Lambda}_t)) + \beta V(f(\hat{\sigma}_t^2, \bar{\Lambda}_t))], \quad (2.15)$$

where $I(\bar{\Lambda}_t)$ is the function defined in (2.12). Thus once we know how to compute the value function for arbitrary values of $\hat{\sigma}_{t+1}^2$, the problem of the optimal choice of a memory structure in any period t can be reduced to the one-period optimization problem stated on the right-hand side of (2.15). This indicates how the memory structure for period t must be chosen to trade off the cost $c(I_t)$ of retaining a more precise memory against the continuation loss $V(\hat{\sigma}_{t+1}^2)$ from having access to a less precise memory in period $t+1$.

Let \mathcal{F} be the class of continuous functions $V(\hat{\sigma}_t^2)$, defined for values of $\hat{\sigma}_t^2$ consistent with (2.4), and consistent with the bounds (2.14) everywhere on this domain. Then (2.15) defines a mapping $\Phi : \mathcal{F} \rightarrow \mathcal{F}$: given any conjectured function $V(\hat{\sigma}_{t+1}^2) \in \mathcal{F}$ that is used to evaluate the right-hand side for any value of $\hat{\sigma}_t^2$, the minimized value of the problem on the right-hand side defines a new function $\tilde{V}(\hat{\sigma}_t^2)$ that must also belong to \mathcal{F} . Condition (2.15) states that the value function that defines the minimum achievable continuation loss must be a fixed point of this mapping: a function such that $V = \Phi(V)$.

We can further show that for any function $V \in \mathcal{F}$, the function $\Phi(V)$ defined by the right-hand side of (2.15) is necessarily a monotonically increasing function.¹⁷ It follows that the fixed point $V(\hat{\sigma}_t^2)$ must be a monotonically increasing function. Moreover, we can restrict the domain of the mapping Φ to the subset \mathcal{F}^* of increasing functions.

This then provides us with an approach to computing the optimal memory structure for a given parameterization of our model. First, we find the value function $V(\hat{\sigma}^2) \in \mathcal{F}^*$ that is a fixed point of the mapping Φ , by iterating Φ to convergence. Then, given the value function, we can numerically solve the minimization problem on the right-hand side of (2.15) to determine the optimal transition matrix $\bar{\Lambda}_t$ in any period, once we know the value of $\hat{\sigma}_t^2$ for that period. Solution of this problem also allows us to determine the value of $\hat{\sigma}_{t+1}^2 = f(\hat{\sigma}_t^2, \bar{\Lambda}_t)$, so that the entire sequence of values $\{\hat{\sigma}_\tau^2\}$ for all $\tau \geq t$ can be determined once we know $\hat{\sigma}_t^2$. Finally, we recall that for the initial period $t=0$, the value of $\hat{\sigma}_0^2$ is given

¹⁷See Appendix E.1 for a proof.

by (2.4); we can thus solve for the entire sequence $\{\hat{\sigma}^2\}$ for all $t \geq 0$ by integrating forward from this initial condition.

Once we have determined the sequence of values $\{\hat{\sigma}_t^2\}$ implied by an optimal memory structure for each period, we can determine the elements of the matrix $X_t = X(\hat{\sigma}_t^2)$ each period, using (2.10). We show in the appendix¹⁸ that the degree of uncertainty at the beginning of any period given the structure of the memory chosen for the previous period is given by

$$\Sigma_{t+1} = \Sigma_0 - X_t \bar{\Lambda}'_t.$$

This in turn allows us to calculate the DM's optimal estimate $\hat{\mu}_t$ each period, as a function of the history of realizations $\{y_\tau\}$ of the external state for all $0 \leq \tau \leq t$ and the history of realizations of the DM's memory noise $\{\tilde{\omega}_{\tau+1}\}$ for all $0 \leq \tau \leq t-1$, using (2.1). The DM's complete vector of forecasts z_t each period is then given by (1.3).

2.4 Optimality of a unidimensional memory state

We can show further that the optimal memory state must have a one-dimensional representation. This simplifies the computational formulation of the optimization problem on the right-hand side of (2.15), and provides further insight into the nature of an optimally imprecise memory. Although the information contained in the cognitive state s_t that is relevant for predicting (at time t) what actions will be desirable for the DM in later periods is two-dimensional (both elements of \bar{s}_t matter, if $\rho > 0$, and except when memory is completely uninformative, these are not perfectly correlated with each other), we find that it is optimal for the DM's memory to include only a noisy record of a single linear combination of the two variables. Moreover, this is true regardless of how small memory costs may be.

There is in fact a fairly simple intuition for the result. Note that in any period t , the Kalman filter (2.1) implies that the optimal estimate of the unknown value of μ will be given by a linear function of elements of the cognitive state of the form

$$\hat{\mu}_t = \zeta_t + \delta'_t \bar{m}_t. \tag{2.16}$$

It follows from this that the only information in the memory state m_t that matters for the estimate $\hat{\mu}_t$ is the single quantity $\delta'_t \bar{m}_t$.

We can establish the optimality of a unidimensional memory in the following way. Consider the optimization problem on the right-hand side of (2.15) in any period t , given the degree of uncertainty $\hat{\sigma}_t^2$ determined by the memory structures chosen in earlier periods. The fact that $V(\hat{\sigma}_{t+1}^2)$ is an increasing function, and that $c(I_t)$ is at least weakly increasing, means that an optimal memory structure must minimize the mutual information I_t given the uncertainty $\hat{\sigma}_{t+1}^2$ that it implies for the following period.¹⁹ Hence the optimal choice for

¹⁸See Appendix D.1 for details of the argument.

¹⁹In the case that $c(I)$ is constant over some interval, reducing I_t need not reduce $c(I_t)$, but it cannot increase it; thus the solution to the problem (2.17) must be among the solutions to the problem on the right-hand side of (2.15), even if it is not a unique solution. In such case, showing that the solution to (2.17) is necessarily a singular matrix suffices to show that we can without any loss impose the further constraint in (2.15) that the matrix $\bar{\Lambda}_t$ must be at most of rank one.

$\bar{\Lambda}_t$ must solve the problem

$$\min_{\bar{\Lambda}_t \in \mathcal{L}(X(\hat{\sigma}_t^2))} I(\bar{\Lambda}_t) \quad \text{s.t.} \quad f(\hat{\sigma}_t^2, \bar{\Lambda}_t) \leq \hat{\sigma}_{t+1}^2, \quad (2.17)$$

for given values of $(\hat{\sigma}_t^2, \hat{\sigma}_{t+1}^2)$. We shall show that whenever $(\hat{\sigma}_t^2, \hat{\sigma}_{t+1}^2)$ are such that the set of matrices satisfying the constraint in (2.17) is non-empty,²⁰ the solution $\bar{\Lambda}_t$ to this problem must be at most of rank one. Thus it must be of the special form

$$\bar{\Lambda}_t = \lambda_t X_t v_t v_t', \quad (2.18)$$

where λ_t is a scalar satisfying $0 \leq \lambda \leq 1$ and v_t is a vector normalized to satisfy $v_t' X_t v_t = 1$. It follows that in each period $\bar{m}_{t+1} = X_t v_t \tilde{m}_{t+1}$, where \tilde{m}_{t+1} is a unidimensional memory state with a law of motion

$$\tilde{m}_{t+1} = \lambda_t v_t' \bar{s}_t + \tilde{\omega}_{t+1}, \quad \tilde{\omega}_{t+1} \sim N(0, \lambda_t(1 - \lambda_t)). \quad (2.19)$$

If $\hat{\sigma}_t^2 = \hat{\sigma}_0^2$, the set $\mathcal{L}(X_0)$ consists only of matrices of the form (2.18), with

$$v_t = \frac{w}{(\Omega + \sigma_y^2)^{1/2}(w'w)^{1/2}}, \quad (2.20)$$

because of (2.11). Hence the asserted result is obviously true in that case. Suppose instead that $\hat{\sigma}_t^2 < \hat{\sigma}_0^2$, and consider any matrix $\bar{\Lambda}_t \in \mathcal{L}(X(\hat{\sigma}_t^2))$ that satisfies the constraint in (2.17). If $\bar{\Lambda}_t$ is not itself of rank one (or lower), we shall show that we can choose an alternative transition matrix of the form (2.18), that is also consistent with the constraint in (2.17), but which achieves a lower value of $I(\bar{\Lambda}_t)$.

Let the alternative transition matrix be given by (2.18), with

$$\lambda_t = \frac{\delta'_{t+1} \bar{\Lambda}_t X_t \bar{\Lambda}_t' \delta_{t+1}}{\delta'_{t+1} X_t \bar{\Lambda}_t' \delta_{t+1}}, \quad v_t = \frac{\bar{\Lambda}_t' \delta_{t+1}}{(\delta'_{t+1} \bar{\Lambda}_t X_t \bar{\Lambda}_t' \delta_{t+1})^{1/2}},$$

where $\delta_{t+1} \equiv e_1 - \gamma_{1,t+1}c$ is the vector introduced in (2.16), and let the matrix $\Sigma_{\bar{\omega},t+1}$ be correspondingly modified in the way specified by (2.9). We show in the appendix²¹ that this specification implies that $0 \leq \lambda_t \leq 1$, so that this alternative matrix also belongs to $\mathcal{L}(X_t)$. Moreover, the new memory structure implies a conditional distribution

$$\delta'_{t+1} \bar{m}_{t+1} | s_t \sim N(\delta'_{t+1} \bar{\Lambda}_t \bar{s}_t, \delta'_{t+1} \Sigma_{\bar{\omega},t+1} \delta_{t+1})$$

that is the same as under the original memory structure. This implies that the optimal estimate $\hat{\mu}_{t+1}$ conditional on the cognitive state s_{t+1} will be the same function of \bar{m}_{t+1} and y_{t+1} in the case of the new memory structure, and that the conditional distribution $\hat{\mu}_{t+1} | s_t, y_{t+1}$ will be the same. It follows that $\hat{\sigma}_{t+1}^2$ will be the same, so that the alternative transition matrix also satisfies the constraint in (2.17).

At the same time, we show in the appendix that the reduction in the complexity of memory cannot increase information costs in any period.²² The new memory structure

²⁰Note that this must be the case if $\hat{\sigma}_{t+1}^2$ is chosen optimally given $\hat{\sigma}_t^2$.

²¹See Appendix E.2 for details of the argument.

²²See Appendix E.2 for details of the argument.

consists effectively of a scalar memory state \tilde{m}_{t+1} in each period, which is a multiple of $d'_{t+1}\tilde{m}_{t+1}$, a particular linear combination of the elements of the memory state under the previous memory structure. Hence the information about \bar{s}_t that is revealed by m_{t+1} under the new memory structure (i.e., that is revealed by \tilde{m}_{t+1}) is also information that was revealed by \tilde{m}_{t+1} under the previous memory structure; thus the value of I_t under the previous memory structure must have been at least as large as under the new memory structure. In fact, the only case in which the mutual information will not be reduced by the proposed modification of the memory structure is if all elements of \tilde{m}_{t+1} were multiples of $d'_{t+1}\tilde{m}_{t+1}$; which is to say, only if $\bar{\Lambda}_t$ were already of the special form (2.18).

We conclude, then, that an optimal memory structure must involve a transition matrix in every period of the special form (2.18), so that the memory state each period can be represented by a scalar quantity \tilde{m}_t . The choice of memory structure can then be reduced to a problem of choosing, in each period $t \geq 0$, a scalar quantity $0 \leq \lambda_t \leq 1$, and the direction of a vector v_t (the length of which will then be chosen each period so as to ensure that $v'_t X_t v_t = 1$); the values chosen for these quantities then determine the law of motion for the unidimensional memory state \tilde{m}_{t+1} , specified by (2.19). This in turn determines the elements of the matrix Σ_{t+1} , and hence the value of the gain coefficient $\gamma_{1,t+1}$ in the Kalman filter formula (2.1) and the value of $\hat{\sigma}_{t+1}^2$, which determines the matrix $X_{t+1} = X(\hat{\sigma}_{t+1}^2)$.

For any value $0 \leq \hat{\sigma}_t^2 < \hat{\sigma}_0^2$, let $\mathcal{V}(\hat{\sigma}_t^2)$ be the set of vectors v_t satisfying $v'_t X(\hat{\sigma}_t^2) v_t = 1$. In the case that $\hat{\sigma}_t^2 = \hat{\sigma}_0^2$, we add the further stipulation that $\mathcal{V}(\hat{\sigma}_0^2)$ consists only of the single vector (2.20). Then given a value for $\hat{\sigma}_t^2$, determined by the memory structures for periods $\tau < t$, the memory structure for period t is specified by a scalar quantity $0 \leq \lambda_t \leq 1$ and a vector $v_t \in \mathcal{V}(\hat{\sigma}_t^2)$. These determine a value for $\hat{\sigma}_{t+1}^2 = f(\hat{\sigma}_t^2, \lambda_t, v_t)$, where now the function f is defined for any values of its arguments satisfying $0 \leq \hat{\sigma}_t^2 \leq \hat{\sigma}_0^2$, $0 \leq \lambda_t \leq 1$, and $v_t \in \mathcal{V}(\hat{\sigma}_t^2)$.

Because of the monotonicity of the value function $V(\hat{\sigma}_{t+1}^2)$, the optimal weight vector v_t in any period must be the one that solves the static optimization problem

$$\bar{f}(\hat{\sigma}_t^2, \lambda_t) \equiv \min_{v_t \in \mathcal{V}(\hat{\sigma}_t^2)} f(\hat{\sigma}_t^2, \lambda_t, v_t). \quad (2.21)$$

In the appendix,²³ we give an explicit algebraic solution for the optimal v_t for any given values $0 \leq \hat{\sigma}_t^2 \leq \hat{\sigma}_0^2$ and $0 < \lambda_t \leq 1$,²⁴ and hence for the function $\bar{f}(\hat{\sigma}_t^2, \lambda_t)$. The latter function is also defined when $\lambda_t = 0$, and easily seen to equal $\bar{f}(\hat{\sigma}_t^2, 0) = \hat{\sigma}_t^2$. Thus we can solve for the dynamics of $\{\hat{\sigma}_t^2\}$ implied by any sequence $\{\lambda_t\}$, by iterating the law of motion

$$\hat{\sigma}_{t+1}^2 = \bar{f}(\hat{\sigma}_t^2, \lambda_t),$$

starting from the initial condition $\hat{\sigma}_0^2$ defined in (2.4).

Moreover, it follows from (2.12) that the mutual information associated with the period t memory structure will be given by

$$I_t = -\frac{1}{2} \log(1 - \lambda_t). \quad (2.22)$$

The Bellman equation (2.15) can therefore be written in the simpler form

$$V(\hat{\sigma}_t^2) = \min_{0 \leq \lambda_t \leq 1} [\alpha \hat{\sigma}_t^2 + c(-(1/2) \log(1 - \lambda_t)) + \beta V(\bar{f}(\hat{\sigma}_t^2, \lambda_t))]. \quad (2.23)$$

²³See Appendix E.3 for details.

²⁴Note that no solution is needed in the case that $\lambda_t = 0$, since in this case v_t is undefined.

3 Features of the Model Solution

Here we provide numerical examples of solutions for an optimal memory structure, under alternative assumptions about both the degree of persistence of the process that must be forecasted and the nature of the information cost function. In reporting our results, it is useful to describe the model solution in terms of scale-invariant quantities — that is, ones that are independent of the value of σ_y , indicating the amplitude of the transitory fluctuations in the external state around its mean. Thus we parameterize the degree of prior uncertainty about μ not in terms a value for Ω (the variance of the prior distribution for μ), but rather by a value for $K \equiv \Omega/\sigma_y^2$ (the variance of the prior distribution for μ/σ_y). We similarly measure the degree of uncertainty about μ conditional on the cognitive state at a given point in time (i.e., after a given amount of experience) not in terms of the value of $\hat{\sigma}_t^2$, but rather by the scaled uncertainty measure $\eta_t \equiv \hat{\sigma}_t^2/\sigma_y^2$.

In terms of this scaled uncertainty measure, an optimal memory structure minimizes the value of

$$\sum_{t=0}^{\infty} \beta^t [\eta_t + \tilde{c}(I_t),]$$

a scaled version of (1.6), where the scaled cost function is defined as $\tilde{c}(I) \equiv c(I)/(\alpha\sigma_y^2)$. (Dividing by α further reduces the numbers of parameters that we need to specify in considering the different possible forms that the optimal memory structure may take, since it is only the relative weights on the two loss terms in the objective (1.6) that matter for the optimal memory structure.) Our scale-invariant model is then completely specified by values for the parameters ρ, β, K and the scaled cost function $\tilde{c}(I)$. In our quantitative analysis, we assume that each “period” of our discrete-time model corresponds to a quarter of a year (the variable to be forecasted is a quarterly time series), and hence set $\beta = 0.99$ (implying a discount rate of 4 percent per annum). We consider a variety of values $0 \leq \rho < 1$ for the assumed degree of serial correlation of the external state, and explore the effects of different assumptions regarding the degree of prior uncertainty and the size of information costs.

3.1 The case of a fixed per-period bound on mutual information

We begin by considering the case in which $\tilde{c}(I) = 0$ for all $I \leq \bar{I}$, but the cost is infinite for any value $I > \bar{I}$. (That is, there is a fixed upper bound on the possible mutual information between s_t and m_{t+1} in each period; but any memory structure consistent with this bound is equally feasible and has the same cost.) Here \bar{I} is some finite positive quantity.

Solution for the optimal memory structure is particularly simple in this case. Because of (2.22), the per-period bound on mutual information can equivalently be written as an upper bound $\lambda_t \leq \bar{\lambda}$, where

$$\bar{\lambda} \equiv 1 - e^{-2\bar{I}} > 0.$$

The optimal memory structure in period t is then characterized by the λ_t that minimizes $\bar{f}(\hat{\sigma}_t^2, \lambda_t)$ subject to this constraint. We show in the appendix²⁵ that the minimizing value of λ_t is necessarily the largest feasible value; hence in the solution to this problem, $\lambda_t = \bar{\lambda}$, the value determined by the per-period information bound.

²⁵See Appendix F.1 for details of the argument.

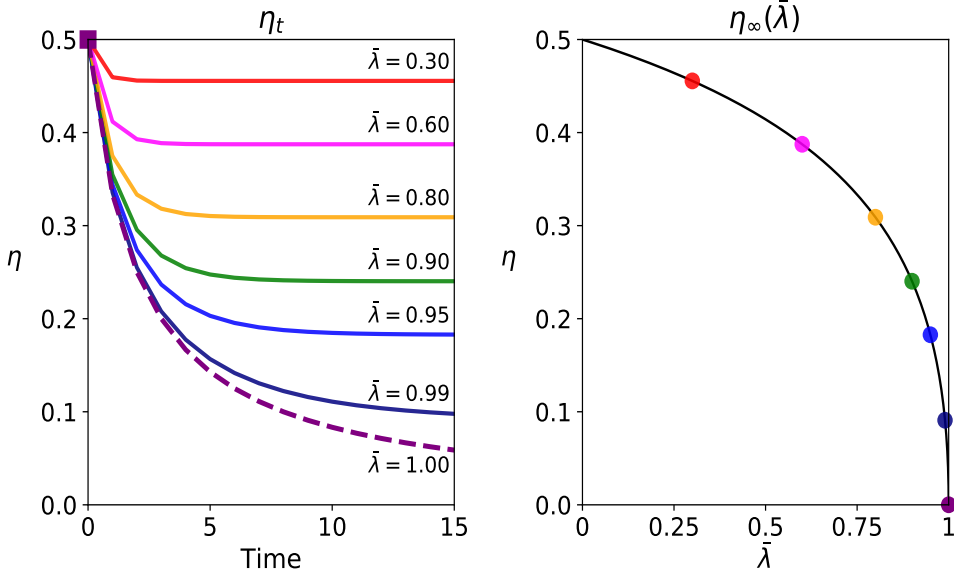


Figure 1: The evolution of scaled uncertainty about μ as the number t of previous (imperfectly remembered) observations grows. The right panel shows the long-run value of scaled uncertainty (to which η_t converges as $t \rightarrow \infty$) as a function of the constraint on the complexity of memory, parameterized by $\bar{\lambda}$.

The dynamics of the uncertainty measure are then given by $\hat{\sigma}_{t+1}^2 = \bar{f}(\hat{\sigma}_t^2, \bar{\lambda})$. In terms of the rescaled variables, the law of motion can be written as

$$\eta_{t+1} = \phi(\eta_t; \bar{\lambda}), \quad (3.1)$$

where $\phi(\eta; \bar{\lambda})$ is a function that is independent of the scale factor σ_y .²⁶

For any value of $\bar{\lambda}$ indicating the tightness of the constraint on the complexity of memory, equation (3.1) indicates how the DM's degree of uncertainty about μ evolves as additional observations of the external state are made. Starting from the initial condition $\eta_0 = K/(K+1)$ implied by (2.4), the law of motion (3.1) can be iterated to obtain a unique solution for the complete sequence of values $\{\eta_t\}$ for all $t \geq 0$. In the limiting case $\bar{\lambda} = 1$ (unlimited memory), the law of motion (3.1) takes the especially simple form

$$\frac{1}{\eta_{t+1}} = \frac{1}{\eta_t} + \frac{1-\rho}{1+\rho}. \quad (3.2)$$

This is simply the standard result for the linear growth in posterior precision under Bayesian updating as additional observations are made; it has the implication that η_t declines monotonically, and converges to zero for large t . Thus in the case of perfect memory, the DM should eventually learn the value of μ with perfect precision, and hence make forecasts of the kind implied by the hypothesis of rational expectations.

When $\bar{\lambda} > 0$, instead, the law of motion (3.1) implies that η_t should decrease initially, as even imprecise memory of the DM's observations of the external state reduces uncertainty

²⁶See Appendix F.1 for an explicit algebraic solution for this function.

to some degree, but that η_t remains bounded away from zero, and converges to a value $\eta_\infty(\bar{\lambda}) > 0$. This is illustrated in Figure 1, which shows the dynamics implied by (3.1) for each of several different values of $\bar{\lambda}$, in the case that $\rho = 0$ and $K = 1$.²⁷ The left panel plots the sequence of values $\{\eta_t\}$ implied by (3.1) for a given value of $\bar{\lambda}$. (Note that the initial value η_0 is the same in each case.) The right panel shows the value η_∞ to which the sequence converges as t grows; this value depends on $\bar{\lambda}$, and the functional relationship between $\bar{\lambda}$ and this limiting degree of uncertainty can be described by a function $\eta_\infty(\bar{\lambda})$, plotted as a smooth curve in the right panel of the figure.

In the case that $\bar{\lambda} = 1$ (shown as a dashed curve in the left panel of Figure 1), the sequence $\{\eta_t\}$ decreases monotonically to zero at the rate predicted by the difference equation (3.2). But for any number of prior observations $t > 0$, the value of η_t remains higher the lower is $\bar{\lambda}$ (that is, the tighter the memory constraint), and the long-run degree of uncertainty about μ , measured by η_∞ , is a decreasing function of $\bar{\lambda}$ as well, as shown by the curve in the right panel of the figure. Because of the limit on the amount of information that can be retained in memory, the DM's uncertainty about the value of μ never falls below a certain level, even in the long run, despite our assumption that the value of μ is fixed for all time. We further observe that the long-run degree of uncertainty η_∞ is larger, the smaller is $\bar{\lambda}$ (that is, the tighter the constraint on memory). In the limit as $\bar{\lambda}$ approaches zero (corresponding to a constraint that memory must be completely uninformative), the long-run degree of uncertainty η_∞ approaches the prior degree of uncertainty $\eta_0 = K/(K + 1)$.

As η_t falls along one of these trajectories, the weight vector v_t that solves the problem (2.21) shifts as well. As η_t converges to the long-run value η_∞ , the optimal weight vector v_t similarly converges to a long-run value v_∞ , indicating the particular linear combination of $\hat{\mu}_t$ and y_t that is imprecisely recorded in memory each period. Associated with this stationary long-run memory structure there will also be a stationary long-run value for the Kalman gain coefficient γ_1 in equation (2.1), and more generally, stationary values for the coefficients of the linear difference equations describing the joint dynamics $\{y_t, \tilde{m}_t\}$ of the external state and the memory state.

These long-run stationary coefficients will depend on the value of $\bar{\lambda}$ (indicating the tightness of the memory constraint) and also on the value of ρ (indicating the degree of persistence of the fluctuations in the external state). Figure 2 indicates how variation in each of these parameters affects several of the long-run stationary coefficients.²⁸ In each panel, a curve shows how the coefficient in question varies as a function of ρ (for values of ρ between 0.0 and 0.9), for a given value of $\bar{\lambda}$; curves of this kind are shown for each of three different values of $\bar{\lambda}$, ranging between $\bar{\lambda} = 0.95$ (in which case memory is relatively precise) and $\bar{\lambda} = 0.30$ (in which case it is much more constrained). All of the curves shown in Figure 2 are again for the case of prior uncertainty $K = 1$.

The upper-right panel of the figure shows the long-run direction vector v_∞ ; the quantity reported on the vertical axis is the long-run value of the ratio v_2/v_1 of the vector's two

²⁷The effects of variation in the parameters ρ and K are illustrated in additional figures shown in Appendix F.1. We use the parameterization $K = 1$ in the figures shown in the text because this value allows a reasonably good fit of the predictions shown in Figure 7 below with the experimental evidence reported by Afrouzi *et al.* (2020).

²⁸See Appendix G.1 for the formulas used to calculate each of the coefficients plotted here as functions of the model parameters.

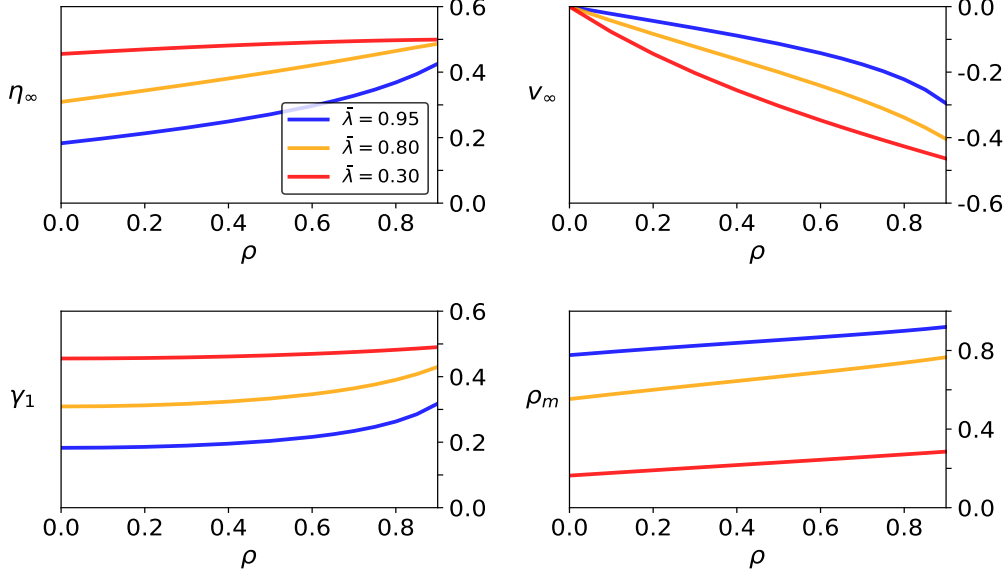


Figure 2: Coefficients describing the optimal memory structure in the long run, as a function of the degree of persistence ρ of the external state, for alternative values of $\bar{\lambda}$. Respective panels show the long-run values for η (measuring uncertainty about μ), the direction vector v (indicating the content of the memory state), the Kalman gain γ_1 (for updating the DM's estimate of μ), and ρ_m (measuring the intrinsic persistence of fluctuations in the memory state).

components.²⁹ Thus a value of -0.3 for this quantity means that the univariate memory state \tilde{m}_{t+1} is (up to a multiplicative factor that does not affect its information content) equal to the value of $\hat{\mu}_t - 0.3y_t$, plus additive Gaussian noise. The figure shows that when $\rho = 0$, the optimal univariate memory state involves $v_2 = 0$; that is, only the current estimate $\hat{\mu}_t$ of the unknown mean is remembered with noise, with the current observation y_t being completely forgotten. This is optimal because when $\rho = 0$, the current value y_t contains no information that is relevant for improving subsequent forecasts of the external state, except to the extent that it helps to improve the DM's estimate of μ (which information is already reflected in the estimate $\hat{\mu}_t$). Instead, when the external state is serially correlated, it is optimal to commit to memory a linear combination of $\hat{\mu}_t$ and the current state y_t ; in the case that $\rho > 0$, the optimal linear combination puts a negative relative weight on y_t , to an extent that is greater the greater the degree of serial correlation, and greater the tighter the constraint on memory.

The upper-left panel of the figure shows the long-run degree of uncertainty about μ , measured by η_∞ . As shown in Figure 1, when $\rho = 0$, η_∞ is a decreasing function of $\bar{\lambda}$. We see in Figure 2 that this is also true when $\rho > 0$. However, for a given memory constraint $\bar{\lambda}$, the long-run value η_∞ is also an increasing function of ρ , with the degree of increase

²⁹This information (together with the value of η_∞ given in the upper left panel) suffices to completely determine the vector v_t , since the vector is normalized so that $v'Xv = 1$. The value of λ (given by the constraint $\bar{\lambda}$), the matrix X (determined by the value of η_∞), and the vector v then completely determine the long-run stationary elements of the matrix $\bar{\Lambda}$ (using (2.18)) and hence also of the matrix $\Sigma_{\bar{\omega}}$ (using (2.9)); thus the dynamics of the memory state given by (2.7) are completely specified.

when the external state is highly persistent being particularly notable when memory is more accurate. The greater the serial correlation of the state, the fewer the effective number of independent noisy observations of μ that the DM receives over any finite time period; thus even under perfect Bayesian updating, equation (3.2) indicates that the rate at which precision is increased by each additional observation is smaller the larger is ρ . In the case of perfect memory, the long-run degree of uncertainty about μ is nonetheless zero (there is simply slower convergence to that long-run value when ρ is large); but with moderately imperfect memory, the effective amount of experience that can ever be drawn upon remains bounded, so that the uncertainty about μ remains larger forever when ρ is larger. When memory is even more imperfect, not much more than one observation (the most recent one) can be used in any event, so that the value of η_∞ is in this case less sensitive to the value of ρ .

In the long run, the dynamics of the cognitive state \bar{s}_t and the memory state \bar{m}_{t+1} are described by linear equations with constant coefficients. The lower-left panel of Figure 2 shows the long-run value for the Kalman gain γ_{1t} in (2.1). With imperfect memory, this is always a quantity between 0 and 1, meaning that a higher value of the current state y_t raises the DM's estimate of the value of μ , though by less than the amount of the increase in y_t . For a given value of ρ , the Kalman gain is larger the tighter the constraint on memory; in the limit as $\bar{\lambda} \rightarrow 1$ (perfect memory), the long-run value of this coefficient approaches zero (as the true value of μ is eventually learned), while in the limit as $\bar{\lambda} \rightarrow 0$ (no memory), the value approaches a maximum value that is still less than one (because of the DM's finite-variance prior).

Finally, the lower-right panel reports the long-run value of ρ_m , a measure of the intrinsic persistence of the memory state. The impulse response function for the effect of a memory-noise innovation $\tilde{\omega}_t$ on the subsequent path of the univariate memory state \tilde{m}_τ is proportional to $(\rho_m)^{\tau-t}$ for all $\tau \geq t$;³⁰ thus the value of ρ_m indicates the rate of exponential decay of the memory state back to its long-run average value. A smaller value of ρ_m means that the contents of memory decay more rapidly; for any value of ρ , the figure shows that ρ_m is smaller, the tighter the memory constraint. At the same time, while a larger value of ρ_m implies that memory persists for a longer time, it also implies that when memory noise creates an erroneous impression of prior experience, this bias in what is recalled about is also corrected more slowly; thus the value of ρ_m is an important determinant of the predicted persistence of forecast bias.

3.2 The case of a linear cost of information

Analysis of the model is more complex when instead the amount of information stored in memory each period can be increased at some finite cost. As an illustration we consider the polar opposite case in which $\tilde{c}(I)$ is a linear function of I , so that the marginal cost of a further increase in the mutual information is independent of how large it already is. Thus we assume that $\tilde{c}(I) = \tilde{\theta} \cdot I$, for some coefficient $\tilde{\theta} > 0$ which parameterizes the cost of memory.

³⁰Here we refer to the difference that the realization of $\tilde{\omega}_t$ makes for the forecasts of \tilde{m}_τ at different horizons $\tau \geq t$, by an observer who knows the true value of μ and the DM's cognitive state at time $t - 1$, in addition to observing the realization of $\tilde{\omega}_t$. See Appendix G.1 for details of the calculation.

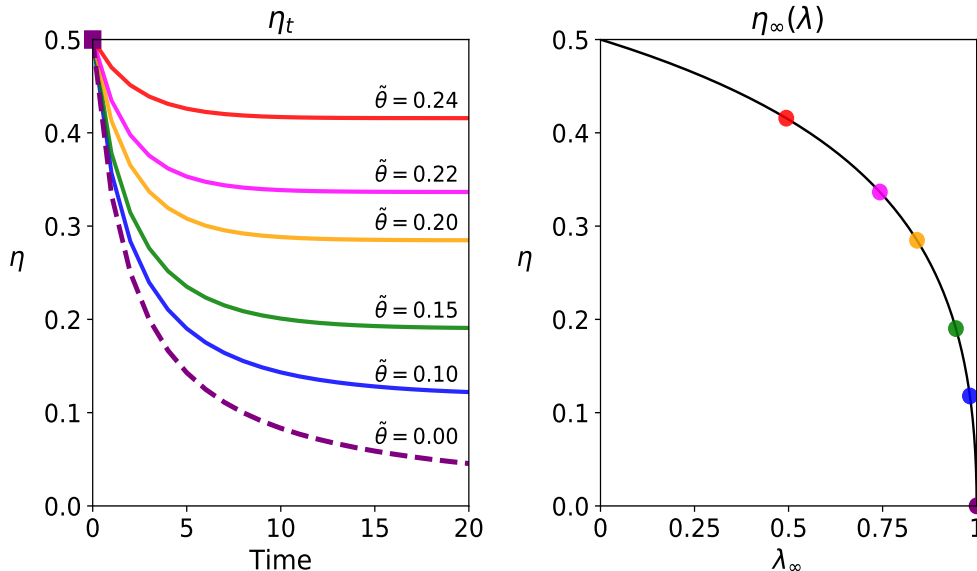


Figure 3: The evolution of scaled uncertainty about μ as the number t of previous (imperfectly remembered) observations grows, now for the case of a linear cost of memory complexity. The right panel shows the long-run value of scaled uncertainty for each value of the cost parameter θ , plotted as a point on the same locus of optimal long-run memory structures as in Figure 1.

In this case, the optimal choice of λ_t in any period will depend on the value of reducing uncertainty in the following period. We note that the value function $V(\hat{\sigma}_{t+1}^2)$ appearing in the Bellman equation (2.23) can be written as $\sigma_y \cdot \tilde{V}(\eta_{t+1})$, where η_{t+1} is the scaled uncertainty measure and the function $\tilde{V}(\eta)$ is independent of the scale factor σ_y (for given values of the parameters K, ρ, β and $\tilde{\theta}$). We can then write the Bellman equation in the scale-invariant form

$$\tilde{V}(\eta_t) = \min_{0 \leq \lambda_t \leq 1} \left\{ \eta_t - \frac{\tilde{\theta}}{2} \log(1 - \lambda_t) + \beta \tilde{V}(\phi(\eta_t; \lambda_t)) \right\}. \quad (3.3)$$

The optimal choice of λ_t in any period will be the value that solves the problem on the right-hand side of (3.3). This problem has a solution $\lambda_t = \lambda^*(\eta_t)$ which depends only on the value of η_t , the degree of uncertainty in period t determined by the memory structures chosen for periods prior to t .

Thus we can solve for the optimal policy function $\lambda^*(\eta_t)$ once we know the value function $\tilde{V}(\eta_{t+1})$, and we can solve numerically for the value function by iterating the Bellman equation (3.3), as discussed further in the appendix.³¹ The policy function $\lambda_t = \lambda^*(\eta_t)$ together with the law of motion

$$\eta_{t+1} = \phi(\eta_t; \lambda_t) \quad (3.4)$$

derived earlier can then be solved for the dynamics of the scaled uncertainty $\{\eta_t\}$ for all $t \geq 0$, starting from the initial condition $\eta_0 = K/(K+1)$.³² The dynamics of scaled uncertainty as

³¹See Appendix F.2 for details.

³²See Appendix F.2 for further discussion of the implied dynamics.

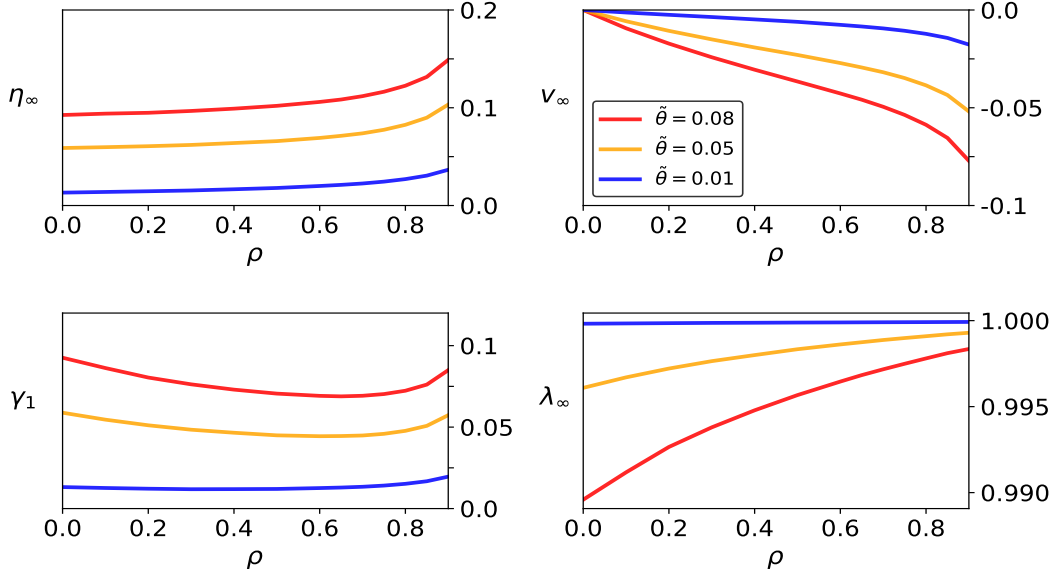


Figure 4: Coefficients describing the optimal memory structure in the long run, as a function of the degree of persistence ρ of the external state, in the case of a linear memory cost function, for alternative values of $\tilde{\theta}$. Respective panels show the long-run values for η , the direction vector v , the Kalman gain γ_1 , and the memory precision coefficient λ .

a function of the number of observations t are shown for progressively larger values of $\tilde{\theta}$ in Figure 3, using the same format as in Figure 1. Once again, we see that while uncertainty about μ eventually falls to zero as a result of when there is no cost of memory complexity, as long as the cost is positive, the value of η_t remains bounded away from zero, and converges asymptotically to a value η_∞ that is higher the higher the cost of memory complexity.

Associated with such an asymptotic degree of uncertainty is a particular long-run memory structure $(\lambda_\infty, v_\infty)$, which will imply a particular long-run value for the Kalman gain γ_1 . The way in which the long-run values of these different quantities vary with different assumptions about the values of ρ and $\tilde{\theta}$ is illustrated in Figure 4. (We use the same convention as in Figure 2 to indicate the direction of the vector v_∞ in the upper-right panel of the figure.) As we vary ρ for a given value of $\tilde{\theta}$, the associated value of λ_∞ changes; hence the fixed- $\tilde{\theta}$ curves shown in Figure 4 do not correspond exactly to any of the fixed- λ curves plotted in Figure 2, even though each of the long-run memory structures associated with a pair $(\rho, \tilde{\theta})$ is identical to the long-run memory structure associated with some pair $(\rho, \bar{\lambda})$. As shown in the lower-right panel of the figure, the optimal λ_∞ rises as ρ increases, for any value of the cost parameter $\tilde{\theta} > 0$; the more persistent the external state that must be forecasted, the more it becomes worthwhile to pay a larger information cost in order to retain a more precise memory of prior experience.

Not surprisingly, we observe that for any value of ρ , increasing the memory cost $\tilde{\theta}$ makes it optimal for the long-run precision of memory λ_∞ to be smaller, and consequently for the long-run degree of uncertainty about μ to be larger. In the case of a sufficiently high value of θ , it will be optimal for memory to be completely uninformative. In fact, this happens for a finite value of $\tilde{\theta}$, and it occurs abruptly, rather than through a gradual increase in the long-run degree of uncertainty η_∞ toward the limiting value of $\eta_0 = K/(K + 1)$ as $\tilde{\theta}$ is

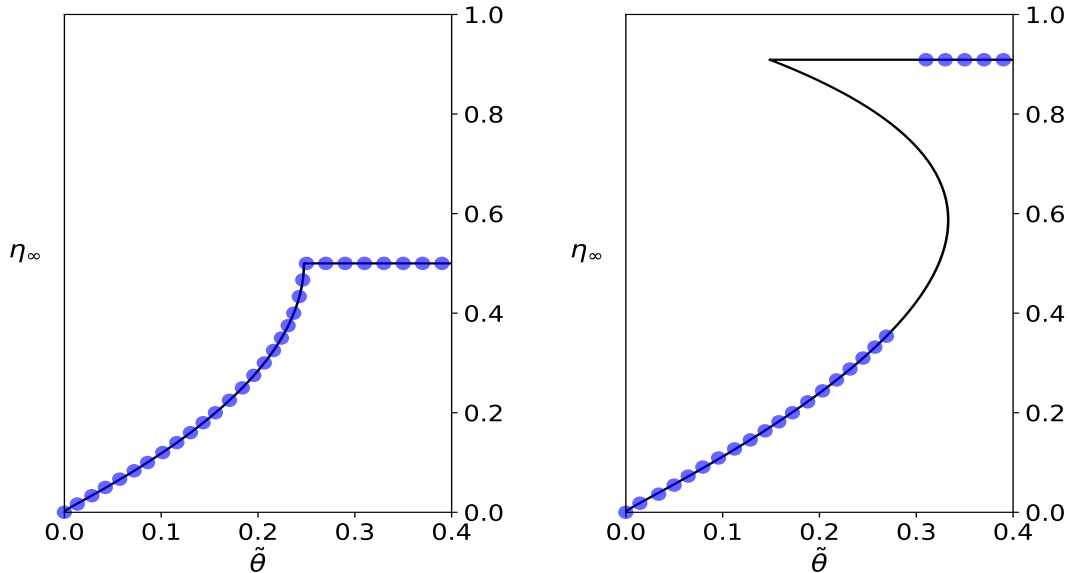


Figure 5: Long-run value of the scaled uncertainty measure η_∞ (blue dots) as a function of the cost parameter $\tilde{\theta}$, in the case of a linear memory cost function. Left panel: $K = 1, \rho = 0$. Right panel: $K = 10, \rho = 0$.

increased. A graph of the relationship between η_∞ and the value of $\tilde{\theta}$ is shown in Figure 5, for the case $\rho = 0$, and two different possible values of K : $K = 1$ and $K = 10$. For each value of $\tilde{\theta}$, the value of η_∞ associated with the optimal memory structure is shown by a large blue dot.

In each panel of this figure, the continuous black curve is the correspondence consisting of all points $(\tilde{\theta}, \eta_\infty)$ such that η_∞ is a stationary solution of the Euler equation associated with the optimization problem on the right-hand side of (3.3).³³ The Euler equation represents a first-order condition for the optimal choice of the degree of precision of memory; satisfaction of this condition is necessary but not sufficient for memory precision leading to $\eta_{t+1} = \eta$ to be optimal starting from a situation in which $\eta_t = \eta$. Because the objective function on the right-hand side of (3.3) is not a convex function, it can have multiple local minima (as well as a local maximum located between two local minima). Which of the local minima represents the global minimum (and hence the optimal memory structure) can jump abruptly as a result of a small change in parameters;³⁴ this is what happens when the value of η_∞ changes abruptly in the right panel of Figure 5, for a value of $\tilde{\theta}$ slightly above 0.28.

In the $K = 10$ case, we see that there need not be a unique value of η_∞ for a given value of $\tilde{\theta}$ that represents a stationary solution to the Euler equation. For any value of $\tilde{\theta}$ greater than a critical value around 0.15, if one starts from $\eta_t = \eta_0$ (a completely uninformative memory), the choice of $\eta_{t+1} = \eta_0$ again represents a local minimum of the objective; hence $\eta = \eta_0$ is a stationary solution of the Euler equation for all of these values of $\tilde{\theta}$, as shown in the figure. However, for values of $\tilde{\theta}$ only moderately larger than the critical value (such as $\tilde{\theta} = 0.20$), this is not the only local minimum, and the global minimum is instead at an

³³See Appendix F.4 for derivation of this equation.

³⁴See Appendix F.3 for a numerical example.

interior choice for λ_t ; this value results in a path $\{\eta_t\}$ that converges to a different stationary value for η_∞ , on the lower branch of the correspondence (as shown for example by the blue dot for $\tilde{\theta} = 0.20$). Yet for values of $\tilde{\theta}$ that exceed a second critical value just above 0.28, the global minimum shifts from the interior minimum to the local minimum at $\eta_{t+1} = \eta_0$. For all values beyond this point, the optimal memory structure involves $\lambda_t = 0$ for all t , so that $\eta_\infty = \eta_0$ (as shown by the blue dots on the upper branch of the correspondence).

Thus while the locus of fixed points $\eta_\infty(\lambda)$ is the same in Figures 1 and 3, all points on this locus represent possible long-run memory structures (attainable through an appropriate choice of $\bar{\lambda}$) in the case of a fixed upper bound on mutual information, but not all of them are always attainable in the case of a linear memory cost function. In the case $K = 1$, the two sets of long-run solutions are identical; but in the case $K = 10$, there is a range of values for η_∞ that are associated with particular (relatively low) values of $\bar{\lambda}$ but do not correspond to any possible value of $\tilde{\theta}$.³⁵

3.3 Stationary fluctuations in the long run

Because our model implies that a DM does not learn the true value of μ with certainty even in the long run, despite an arbitrarily long sequence of observations of the external state, over which time the coefficients of the data-generating process (1.1) are assumed not to change, it follows that the DM’s forecasts can be quite different from rational-expectations forecasts — that is, the forecasts of an ideal statistician who knows the true coefficient values. From the standpoint of an observer who is able to determine the true process, the forecasts of the DM with limited memory will appear to be systematically biased. The biases in the DM’s forecasts will furthermore fluctuate over time, in response both to variations in the external state (to which the DM reacts differently than someone with rational expectations would) and to noise in the evolution of the memory state.

We obtain a particularly simple characterization of the systematic pattern of forecast biases if we consider the long run — the predictions of the equations in the previous two sections in the case of very large values of t , so that η_t has converged to the constant value η_∞ , λ_t has converged to λ_∞ , and so on. In this case, our model, like the model of “natural expectations” of Fuster *et al.* (2010, 2011), predicts a stationary pattern of forecast biases that do not reflect incomplete adjustment to a new environment.

In the long run, equations (1.1), (2.1), and (2.7) become a system of linear equations with constant coefficients and Gaussian innovation terms, describing the evolution of the DM’s cognitive state. This system of equations can be reduced to a VAR(1) system

$$\tilde{s}_{t+1} = f\mu + F\tilde{s}_t + u_{t+1}, \quad u_{t+1} \sim N(0, \Sigma_u) \quad (3.5)$$

where

$$\tilde{s}_t \equiv \begin{bmatrix} \tilde{m}_t \\ y_t \end{bmatrix}, \quad u_{t+1} \equiv \begin{bmatrix} \tilde{\omega}_{t+1} \\ \epsilon_{y,t+1} \end{bmatrix},$$

³⁵We can show analytically that the continuous relationship shown in the left panel of Figure 5 occurs for all $K \leq 1$ when $\rho = 0$, while the backward-bending correspondence and consequent discontinuous relationship between $\tilde{\theta}$ and η_∞ occurs for all $K > 1$. See Appendix F.4 for further explanation.

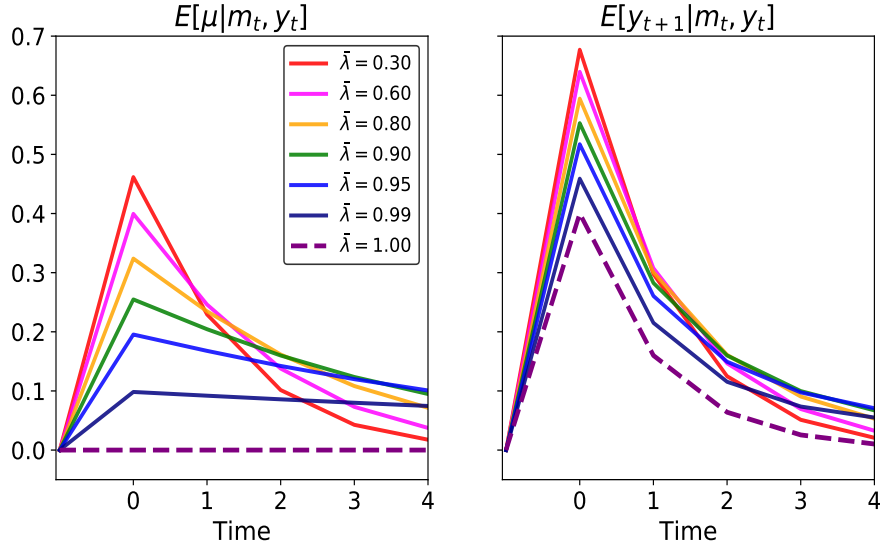


Figure 6: Impulse responses of the DM’s estimate of μ (left panel) and one-period-ahead forecast of the state (right panel) to a unit positive innovation in the observed value of y_t at the time marked as “time = 0” on the horizontal axis. Responses are plotted for alternative values of the information bound $\bar{\lambda}$, in the case that $K = 1, \rho = 0.4$.

and f, F and Σ_u are a 2-vector and two 2×2 matrices of constant coefficients respectively. In this vector system, the first equation is obtained by substituting (2.1) into (2.19), while the second equation is given by (1.1).

The matrix F furthermore has an upper-triangular form, while Σ_u is diagonal. We show in the appendix that the eigenvalues of the matrix F are ρ and ρ_m .³⁶ We further show that $0 < \rho_m < 1$, so that both y_t and \tilde{m}_t exhibit stationary fluctuations around well-defined long-run average values which depend linearly on μ . The two independent exogenous sources of variation in this system are the innovations $\epsilon_{y,t+1}$ in the external state and the memory noise innovations $\tilde{\omega}_{t+1}$.

The DM’s optimal estimate of μ at each point in time, $\hat{\mu}_t$, as well as her optimal forecast of the external state at any horizon $\tau > t$,

$$\hat{y}_{\tau|t} = E[y_{\tau} | \tilde{m}_t, y_t] = (1 - \rho^{\tau-t})\hat{\mu}_t + \rho^{\tau-t}y_t, \quad (3.6)$$

will then be linear functions of the elements of \tilde{s}_t , with coefficients that are also time-invariant. We thus obtain a stationary multivariate Gaussian distribution for any number of leads and lags of the external state, the DM’s memory state, and the DM’s estimates and forecasts. This allows us to analyze not only the extent to which the DM’s forecasts should differ from rational-expectations forecasts, but the correlation that one should observe between the bias in the DM’s forecasts and other observable variables.

In particular, the biases in the DM’s forecasts will be correlated with the evolution of the external state. An unexpectedly high observed value for y_t will be interpreted (because of the DM’s uncertainty about μ) as implying a higher optimal estimate of μ , and this increase

³⁶See Appendix G.1 for the derivation.

in the DM's estimate of μ will furthermore persist, decaying only gradually in subsequent periods. This is illustrated in the left panel of Figure 6, which shows the impulse response function for $\hat{\mu}_\tau$ to a unit positive innovation in the value of y_t . The response is plotted for a variety of alternative values for the information bound $\bar{\lambda}$, in the case that $K = 1$ and $\rho = 0.4$.

In the case that $\bar{\lambda} = 1$ (perfect memory), the value of μ is learned with perfect precision, and as a consequence there is no effect (in the long run, depicted here) of fluctuations in y_t on the DM's estimate of μ . (The Kalman gain γ_1 has a long-run value of zero in this case.) Instead, for values of $\bar{\lambda} < 1$, a higher observed value of y_t leads the DM to increase her estimate $\hat{\mu}_t$ (the Kalman gain is positive). The estimate $\hat{\mu}_\tau$ remains higher (on average) in subsequent periods as well. The memory state \tilde{m}_{t+1} carried into the period following the innovation is a noisy record of $\hat{\mu}_t$, and hence is higher because of the increase in y_t ; this increases the average value of the estimate $\hat{\mu}_{t+1}$, which increases the average value of the memory state \tilde{m}_{t+2} , and so on. The tighter the memory constraint (the lower the value of $\bar{\lambda}$), the greater the effect of the innovation in y_t on $\hat{\mu}_t$, because the DM is more uncertain about the value of μ before observing y_t ; however, the effect on the DM's estimate of μ is also more transient the lower the value of $\bar{\lambda}$, because less information is retained from one period to the next about past cognitive states.

These effects on the DM's optimal estimate of μ then feed into her optimal forecast of the external state at any future horizon τ , because of (3.6). As an illustration, the right panel of Figure 6 shows the impulse response of the one-quarter-ahead forecast $\hat{y}_{\tau+1|\tau}$ to a unit positive innovation in y_t , using the same conventions as in the left panel.³⁸ When $\rho > 0$, the rational-expectations forecast (corresponding to $\bar{\lambda} = 1$ in the figure) is itself increased by a positive innovation in y_t (by an amount equal to fraction ρ of the innovation), and the increase in the forecast is furthermore persistent (decaying back to its original level at a rate proportional to $\rho^{\tau-t}$). But when $\bar{\lambda} < 1$, the forecast is increased by even more, owing to the fact that the higher observation of y_t increases the DM's estimate of μ as well. This additional effect on the forecast is initially larger the smaller is $\bar{\lambda}$; but a smaller $\bar{\lambda}$ (tighter memory constraint) also causes the additional effect to die out more rapidly, since its propagation can only be through the DM's memory of her previous judgment about the value of μ .

Thus our model predicts that forecasts of the future value of a variable will over-react to news about the current value of that variable (assuming, as is often the case with economic time series, that the variable in question exhibits positive serial correlation). Positive serial correlation means that a higher current observation should increase somewhat one's forecast of the variable's future value, even under rational expectations; but imperfect memory results in a larger increase in the forecast than is consistent with rational expectations. The model also predicts that biases of this kind will persist for some time. Once a situation occurs that leads the DM to over-estimate the future level of some time series, the DM will as a consequence continue (on average) to over-estimate the future level of that variable for several more quarters.

³⁷See Appendix G.1 for illustration of how this figure would change under alternative assumptions about the degree of persistence of the fluctuations in the external state.

³⁸The corresponding impulse responses for alternative values of ρ are again shown in Appendix G.1.

3.4 “Recency bias” in expectation formation

One type of systematic difference between observed expectations and those of a perfect Bayesian decision maker that has often been reported is “recency bias” (e.g., Malmendier *et al.*, 2017) — a tendency for expectations to be influenced more by more recent observations, even when in principle, observations of a given time series at earlier dates should be equally relevant as a basis for inference. Our model predicts that such a bias should exist, as a consequence of optimal adaptation to limited memory precision (or to the cost of maintaining a more precise memory). Observations of the external state farther in the past are recalled with more noise, and as a consequence are given less weight in estimating parameters of the data generating process than would be optimal in the case of a perfect memory of past data.

The system (3.5) implies that, in the case that data have been generated in accordance with this law of motion for a sufficiently long time, we can express the value of the memory state \tilde{m}_{t+1} as a function of the sequence of external states $\{y_\tau\}$ for $\tau \leq t$ and the sequence of memory noise realizations $\{\tilde{\omega}_{\tau+1}\}$ for $\tau \leq t$:

$$\tilde{m}_{t+1} = F_{12} \cdot \sum_{j=0}^{\infty} (\rho_m)^j y_{t-j} + \tilde{\omega}_{t+1}^{sum}, \quad (3.7)$$

where F_{12} is the (1, 2) element of the matrix F in (3.5) and

$$\tilde{\omega}_{t+1}^{sum} \equiv \sum_{j=0}^{\infty} (\rho_m)^j \tilde{\omega}_{t+1-j} \quad (3.8)$$

is a serially correlated Gaussian noise term.³⁹

Equation (2.1) implies that a DM’s estimate of the unknown mean μ of the external state is given by a linear relation of the form

$$\hat{\mu}_t = \xi \tilde{m}_t + \gamma_1 y_t, \quad (3.9)$$

where the coefficient $\xi > 0$ is defined in the appendix. Using (3.7) to substitute for the memory state in this expression, we see that we can write the estimate in the form

$$\hat{\mu}_t = \sum_{j=0}^{\infty} \alpha_j y_{t-j} + \xi \tilde{\omega}_t^{sum}, \quad (3.10)$$

where the weights $\{\alpha_j\}$ are all positive, and the weights for $j \geq 1$ decrease exponentially: $\alpha_j = \alpha_1 (\rho_m)^{j-1}$.

The forecasts specified by (3.6) using this value for $\hat{\mu}_t$ are similar to those implied by a model of least-squares learning (Evans and Honkapohja, 2001) in which the DM is assumed to know that the variable’s law of motion is of the form (1.1); the value of the coefficient ρ is assumed to be known while μ must be estimated; and the unknown coefficient is estimated using a “constant-gain” estimator.⁴⁰ The biases in forecasts predicted by our model will

³⁹This is a stationary random process with a finite unconditional variance, since $0 < \rho_m < 1$ as shown in Appendix G.1.

⁴⁰The differences between (3.10) and a standard constant-gain estimate of the mean of a series are the fact that the coefficient α_0 is differently specified, and the presence of the Gaussian error term.

therefore have important similarities to those of a model of constant-gain learning, of the kind included in estimated macroeconomic models by authors such as Milani (2007, 2014) and Slobodyan and Wouters (2012).

We provide, however, a justification for the declining weight on observations farther in the past, as a consequence of optimal forecasting based on an imperfect memory, and furthermore endogenize the nature of that memory. The fact that our model predicts decreasing weights on observations made farther in the past is a notable difference between our model and the one proposed by Afrouzi *et al.* (2020).

4 Predictable Forecast Errors

An important consequence of optimal Bayesian inference with perfect memory (as assumed under the hypothesis of rational expectations) is that the error in a forecast should not itself be forecastable on the basis of any information available to the forecaster, at or before the time of the forecast in question. Thus if we let $\hat{y}_{t+h|t}$ denote a DM’s forecast at time t of the value of the external state at time $t+h$, the forecast error⁴¹ $FE_t \equiv y_{t+h} - \hat{y}_{t+h|t}$ should be uncorrelated with any variable z_t the value of which is known to the DM at time t (or earlier), either because it has been publicly observable or because it is part of the DM’s own cognitive state. Many econometric investigations of the consistency of observed forecasts with the hypothesis of rational expectations have accordingly been based on regressions of FE_t on other variables that ought to be known to the forecaster, testing the null hypothesis that all such regression coefficients should equal zero. Here we discuss the extent to which our model can account for some widely discussed examples of evidence against this null hypothesis; we particularly discuss evidence indicating over-reaction of subjective expectations to news about the series that is to be forecasted.

4.1 Evidence of over-reaction: the response of forecasts to fluctuations in the state

As noted in the introduction, Afrouzi *et al.* (2020) conduct a laboratory experiment in which forecasts of a stationary AR(1) process are elicited from subjects. They find that subjects’ expectations over-react to innovations in this process, as predicted by our model (as well as the related model of noisy memory that they discuss). They give particular emphasis to a measure of over-reaction in which a subject’s forecast $\hat{y}_{t+h|t}$ (where h is the number of realizations in advance for which the forecast is solicited in trial t) is regressed on the realization of the variable just before the forecast is solicited:

$$\hat{y}_{t+h|t} = \alpha_h^{subj} + \rho_h^{subj} y_t + v_t. \quad (4.1)$$

A separate regression (with coefficients α_h, ρ_h^{subj}) can be estimated for each of several horizons h . Afrouzi *et al.* are interested in the difference between the “subjective degree of

⁴¹Note that we give this variable a time subscript t to indicate that it is the error in the forecast made at time t ; the value of the random variable FE_t is not revealed however until date $t+h$.

persistence” measured by the estimated coefficient ρ_h^{subj} and the corresponding coefficient ρ_h in a regression using actual outcomes:

$$y_{t+h} = \alpha_h + \rho_h y_t + u_{t+h}. \quad (4.2)$$

The authors measure the degree of over-reaction of expectations to news by the extent to which ρ_h^{subj} is larger than ρ^h . Note that this is an example of a test of the predictability of forecast errors, since the coefficient of a regression of FE_t on y_t will equal $\rho^h - \rho_h^{subj}$.

We can investigate what our model of expectation formation on the basis of an imperfect memory implies about the relationship between ρ_h^{subj} and ρ_h in the case of a stationary AR(1) process. Here we consider the predicted values of the regression coefficients in the long run, as the length of the time series used to estimate them goes to infinity. The law of motion (1.1) implies that for any horizon $h \geq 1$, the joint distribution of y_t and y_{t+h} (conditional on the value of μ) will be bivariate Gaussian, with

$$E[y_{t+h} | \mu, y_t] = (1 - \rho^h)\mu + \rho^h y_t.$$

Hence with a sufficiently long series of observations, the coefficients in a regression of the form (4.2) should approach the asymptotic values

$$\alpha_h = (1 - \rho^h)\mu, \quad \rho_h = \rho^h.$$

(Here we assume that the regression uses an arbitrarily long sequence of realizations of a process for which there is a single, unchanging value of μ .)

Equation (3.6) implies that subjective forecasts should be given by

$$\hat{y}_{t+h|t} = (1 - \rho^h)\hat{\mu}_t + \rho^h y_t,$$

so that the predicted coefficient ρ_h^{subj} in regression (4.1) will equal

$$\rho_h^{subj} = (1 - \rho^h)\beta_{\hat{\mu}|y} + \rho^h = (1 - \rho_h)\beta_{\hat{\mu}|y} + \rho_h, \quad (4.3)$$

where $\beta_{\hat{\mu}|y}$ is the coefficient in a regression of $\hat{\mu}_t$ on y_t ,

$$\beta_{\hat{\mu}|y} = \frac{\text{cov}[\hat{\mu}_t, y_t | \mu]}{\text{var}[y_t | \mu]} = \frac{\text{cov}[\hat{\mu}_t, y_t | \mu]}{\sigma_y^2}.$$

We show in the appendix how to calculate this coefficient as a function of the model parameters.⁴²

Importantly, our numerical solutions indicate that $\hat{\mu}_t$ and y_t are always positively correlated (conditional on μ). This is because a positive innovation in the external state y_t raises (or at least never lowers) the expected value of y_τ for all $\tau \geq t$, and at the same time also raises the expected value of $\hat{\mu}_\tau$ for all $\tau \geq t$ (as illustrated in Figure 6 and similar figures in the appendix). Since the memory noise has no effect on the evolution of the external state, there are no shocks that move $\hat{\mu}_t$ and y_t in opposite directions, while some (at least the innovation ϵ_{yt}) move both of them in the same direction. But given that $\beta_{\hat{\mu}|y} > 0$, equation

⁴²See Appendix G.3 for details.

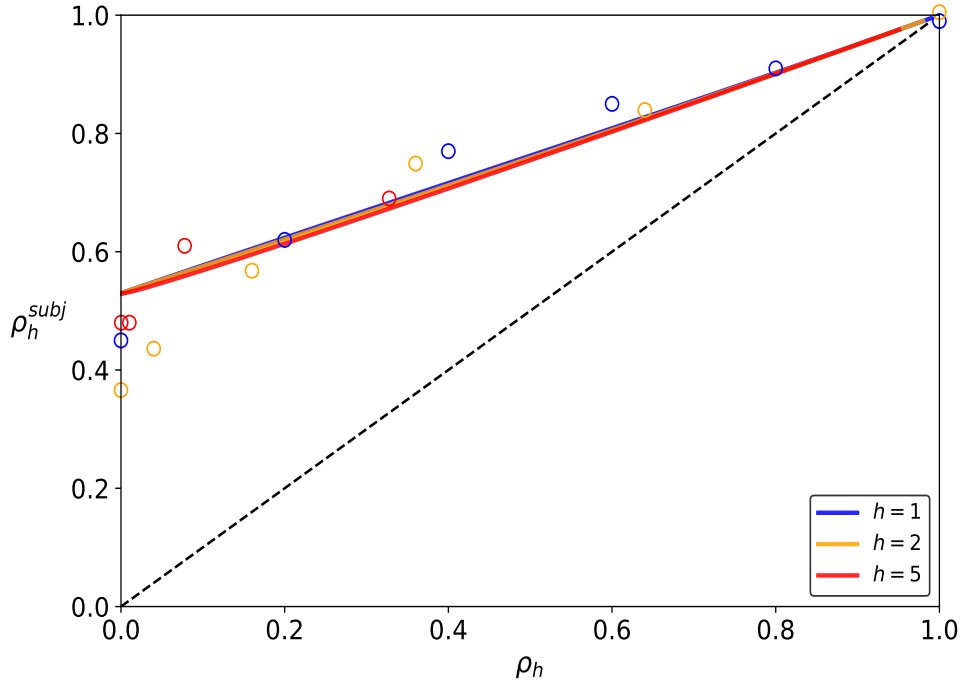


Figure 7: Comparison of the values for the regression coefficients ρ_h and ρ_h^{subj} for different values of ρ and h . (The figure is shown for the case $K = 1, \bar{\lambda} = 0.3$.) The diagonal line indicates the prediction of the rational-expectations hypothesis.

(4.3) implies that $\rho_h^{subj} > \rho_h$; that is, our model implies over-reaction of the kind exhibited by the forecasts of the subjects of Afrouzi *et al.*

Equation (4.3) also implies that for fixed values of the model parameters other than ρ , the over-reaction measure $\rho_h^{subj} - \rho_h$ converges to zero as $\rho \rightarrow 1$, for any forecast horizon h .⁴³ This is also approximately true of the regression coefficients reported by Afrouzi *et al.* (see their Figures 2B, 5A, and 5B). Indeed, these authors stress the finding that in their data, the discrepancy $\rho_h^{subj} - \rho_h$ is much larger when ρ_h is relatively small (either because ρ is small, or because ρ is well below one and h is large). This is also true in numerical solutions of our model as indicated in Figure 7.

One of the more striking features of the regressions reported by Afrouzi *et al.* is that ρ_h^{subj} is well approximated by an increasing function of ρ_h , with approximately the same functional relationship regardless of whether the variation in ρ_h occurs as a result of variation in ρ or variation in h .⁴⁴ The relationship $\rho^{subj}(\rho)$ is furthermore an upward-sloping one, with a slope much less than one, starting well above the diagonal for low values of ρ and approaching the

⁴³This prediction depends on $\beta_{\hat{\mu}|y}$ remaining bounded as ρ approaches 1. This is the case in our numerical solutions, both when $\bar{\lambda}$ is held constant as ρ is varied (as in Figure 2) and when $\bar{\theta}$ is held constant as ρ is varied (as in Figure 4).

⁴⁴This was shown in an earlier version of the paper now circulated as Afrouzi *et al.* (2020), though this figure is omitted from their most recent draft.

diagonal as $\rho \rightarrow 1$. (See the plot of their regression coefficients in Figure 7.⁴⁵) While our model does not imply that a functional relationship of that kind should hold precisely, it is worth noting that to the extent that the value of $\beta_{\hat{\mu}|y}$ remains approximately the same as one varies ρ , (4.3) implies that the value of ρ_h^{subj} should be nearly the same for all pairs (ρ, h) that imply the same value of ρ_h . Perhaps more to the point, our model can be parameterized so that it simultaneously fits the experimental evidence for each of the three different horizons for which forecasts are solicited in the experiment of Afrouzi *et al.*

Figure 7 plots the predicted value of ρ_h^{subj} against the value of ρ_h , for each of several different horizons h , each represented by a distinct curve; the curves are shown for the case in which $K = 1$ and $\bar{\lambda} = 0.3$. Along each curve, the variation in ρ_h is due purely to variation in ρ . (The fact that $\bar{\lambda}$ is fixed despite variation in ρ means that we assume a fixed upper bound on the mutual information, as in section 3.1, rather than a convex cost function.) The horizons used are $h = 1, 2$ and 5 , as these are the horizons for which Afrouzi *et al.* elicit forecasts from their subjects; the regression coefficients that they estimate for various combinations of ρ and h are indicated by the circles in the figure (with colors indicating the horizon h).

The three curves are not exactly the same, since in our model $\beta_{\hat{\mu}|y}$ is a function of ρ (but the same for all values of h), rather than being a function only of ρ_h . Nonetheless, for the parameterization chosen here, $\beta_{\hat{\mu}|y}$ is nearly constant as ρ is varied; as a consequence, the relationship between ρ_h and ρ_h^{subj} predicted by (4.3) is close to a linear one, and is nearly the same for all values of h . Our model therefore provides quite a good account of the effects of variation in either ρ or h on the value of ρ_h^{subj} , as indicated by the fact that none of the circles in Figure 7 are far from the corresponding curve.

4.2 Evidence of over-reaction: forecast revisions and forecast error

A comparison of the coefficient ρ_h^{subj} with ρ_h provides a fairly straightforward test of over-reaction of forecasts to news about the variable being forecasted; but the null hypothesis that ρ_h^{subj} should equal ρ_h only follows from rational expectations in the case that one is sure that forecasters at time t have observed the external state y_t , and calculation of the predicted value ρ_h requires knowledge of the true data-generating process. Both of these assumptions make sense in the case of the laboratory experiment of Afrouzi *et al.* (2020); but they are more debatable in the case of forecasts of economic time series outside laboratory settings.

Bordalo *et al.* (2020a) use a different approach to document systematic departures from rational expectations in surveys of professional forecasters. Following a proposal by Coibion and Gorodnichenko (2015), they regress the error (that eventually becomes known) in a given forecaster’s forecast of a future data release on the revision that the forecast represents, relative to the same forecaster’s forecast of the same future variable at an earlier time. That is, they test whether the coefficient b is different from zero in a regression specification of the form

$$FE_t = a + b \cdot FR_t + u_t, \tag{4.4}$$

⁴⁵The data plotted here are based on Figures 2B, 5A, and 5B of Afrouzi *et al.* (2020).

where FE_t is the error (as defined above) in the forecast at time t for some horizon $h > 0$, and

$$FR_t \equiv \hat{y}_{t+h|t} - \hat{y}_{t+h|t-1}$$

is the revision of this forecast between time $t - 1$ and time t .

If the forecast $\hat{y}_{t+h|t}$ represents the correctly calculated expectation of y_{t+h} conditional on the forecaster's information set at time t , then the forecast error FE_t should not be forecastable on the basis of any information available to the forecaster at time t , as discussed above. Bordalo *et al.* point out that even if one is agnostic about which external developments are observed by forecasters (or how accurately they observe them), as long as one supposes that the forecaster's *own past cognitive states* are known with complete precision, then the size of the forecast revision FR_t should be part of the information set; hence the coefficient b in (4.4) should equal zero under a hypothesis of correct Bayesian inference from the forecaster's information set. A coefficient $b \neq 0$ allows one to reject not just the full-information rational-expectations hypothesis, but also models that assume that people's choices are optimal Bayesian responses conditional upon a noisy cognitive state (but with perfect memory), such as the models of Sims (2003) or Woodford (2003).

Bordalo *et al.* (2020a) find instead that b is significantly negative in the case of many macroeconomic and financial time series. (Afrouzi *et al.*, 2020, find that the same is true of the forecasts elicited in their experiment.⁴⁶) Bordalo *et al.* interpret this negative sign as evidence of over-reaction of forecasts to economic news arriving between the dates of the two successive forecasts: news that implies that one's previous forecast was too low results in an upward revision that is too large, so that the occurrence of an upward revision is correlated with the second forecast turning out to be too high.

As discussed above, our model implies that there will be over-reaction to new realizations of the external state, and indeed our model predicts that one can easily have a negative coefficient b in a regression of the form (4.4). The theoretically predicted asymptotic value for the coefficient b in the case of a long enough series of observations from an environment with unchanging statistics (including a fixed value of μ) is given by

$$b = \frac{\text{cov}[FE_t, FR_t | \mu]}{\text{var}[FR_t | \mu]},$$

where the forecast error and forecast revision variables are defined above. Since the denominator is necessarily positive (in any case in which the forecast is not constant at all times), the coefficient b should be negative if and only the covariance between FE_t and FR_t is negative.

That this can easily be the case can be illustrated by considering the simple case of an i.i.d. process ($\rho = 0$) for the external state. In this case, (3.6) implies that $\hat{y}_{t+h|t} = \hat{\mu}_t$, for any forecast horizon $h \geq 1$. In this case we have

$$\text{var}[FR_t | \mu] = \text{var}[\hat{\mu}_t - \hat{\mu}_{t-1} | \mu] = 2(1 - \rho_{\hat{\mu}})\text{var}[\hat{\mu}_t | \mu],$$

⁴⁶See Figure 2A in their paper. Indeed, the evidence for $b < 0$ as a general regularity is even stronger in the laboratory data of Afrouzi *et al.* than in the field data of Bordalo *et al.*

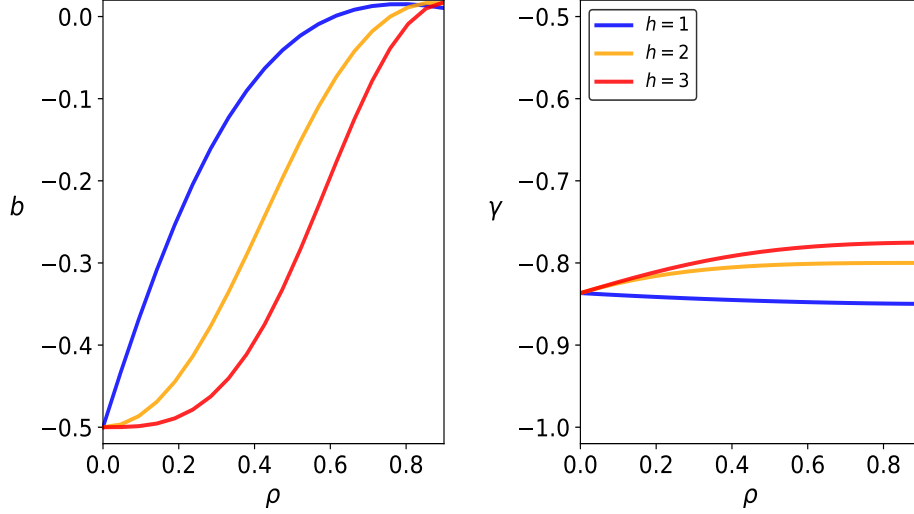


Figure 8: Predicted value of the coefficient b from a regression of forecast errors on the size of the revision of the forecast (left panel), and the coefficient γ from a regression of forecast revisions on the deviation of an individual’s forecast from the consensus forecast (right panel). In each case, results are shown for external state processes of different degrees of serial correlation ρ . As in Figure 7, we assume that $K = 1$, $\bar{\lambda} = 0.3$.

where $\rho_{\hat{\mu}}$ is the coefficient of serial correlation of the stationary fluctuations in $\hat{\mu}_t$, and

$$\begin{aligned} \text{cov}[FE_t, FR_t | \mu] &= \text{cov}[y_{t+h} - \hat{\mu}_t, \hat{\mu}_t - \hat{\mu}_{t-1} | \mu] \\ &= \text{cov}[-\hat{\mu}_t, \hat{\mu}_t - \hat{\mu}_{t-1} | \mu] = -(1 - \rho_{\hat{\mu}})\text{var}[\hat{\mu}_t | \mu]. \end{aligned}$$

(Here the first expression on the second line follows from the fact that y_{t+h} is completely uncorrelated with any variables observable at time t or earlier, conditional on the value of μ , when $\rho = 0$.) Hence in this case we obtain the prediction $b = -1/2$, simply as a consequence of the fact that our model implies stationary fluctuations in $\hat{\mu}_t$ around some long-run average estimate, for any parameter values with $\lambda_\infty < 1$.

Our numerical solutions indicate that b is often negative in the case of positive serial correlation in the external state as well, as illustrated in the left panel of Figure 8.⁴⁷ The figure shows the predicted value of the coefficient b as the coefficient of serial correlation ρ varies between 0 and 1. The figure is computed under the assumption that $K = 1$ and $\bar{\lambda} = 0.3$, as in Figure 7. We see that the predicted degree of over-reaction (as measured by the degree to which $b < 0$) is greatest when $\rho = 0$; the coefficient equals -0.5 when $\rho = 0$ (as explained in the previous paragraph), but is less negative when $\rho > 0$, and near zero for large values of ρ . This is also what Afrouzi *et al.* find to be true of the forecasts elicited in their laboratory experiment.⁴⁸

A similar regularity is observed in the case of professional forecasts of the economic time series considered by Bordalo *et al.* (2020a). The estimates that they obtain for b mainly

⁴⁷Formulas that can be used to calculate b are given in Appendix G.3.

⁴⁸Again, see their Figure 2A.

fall in or near the interval $[-0.5, 0]$. Moreover, in the case of the highly persistent series that they consider, the value of b is around zero on average (sometimes slightly negative, but sometimes slightly positive); in the case of the series that they consider with a coefficient of serial correlation less than 0.1, the value of b is nearly as negative as -0.5; and for the series that they consider with intermediate degrees of serial correlation, b is negative but much less negative than -0.5.⁴⁹ Our model is not only able to explain why negative values of b are often obtained, but also why these are almost always between small positive values and -0.5, and why the coefficient is more negative for the least persistent time series.

4.3 Idiosyncratic noise in individual forecasts

Another kind of evidence of systematic bias in the forecasts announced by individual forecasters is provided by Fuhrer (2018). Fuhrer shows that subsequent revisions of the forecasts of an individual forecaster are partially forecastable on the basis of information available (at least to the community of forecasters in general) at the time of the original forecast, a result inconsistent with the hypothesis of full-information rational expectations (under which not only should all forecasters be ideal Bayesian statisticians, but all should share a common information set).

Suppose now that we let $\hat{y}_{t+h|t}^i$ be the forecast of y_{t+h} at time t by forecaster i , while $\hat{y}_{t+h|t}^{cons}$ is the “consensus forecast,” the median of the forecasts at time t made by the different forecasters in a given survey. Fuhrer reports regressions of the form

$$\hat{y}_{t+h|t}^i - \hat{y}_{t+h|t-1}^i = a + \gamma \cdot (\hat{y}_{t+h|t-1}^i - \hat{y}_{t+h|t-1}^{cons}) + e'Z_t^i + u_t, \quad (4.5)$$

where Z_t^i is a vector of other forecaster-specific control variables (that vary across specifications). His main finding, obtained for forecasts of several different aggregate variables, and robust both to different choices for the horizon h and the control variables included, is that the coefficient γ is found to be significantly negative (for example, between -0.5 and -0.6 in the case of revisions of inflation forecasts by members of the Survey of Professional Forecasters, where the forecasts are collected at a quarterly frequency). This indicates a tendency of forecasters to subsequently revise their forecasts so as to partially eliminate the previous gap between their forecast and the consensus forecast.

Many models of biased expectation formation proposed in the previous literature predict systematic departures from rational expectations, and hence allow forecast revisions to be predictable by information that was publicly available at the earlier date; but they nonetheless do not predict that the variable considered by Fuhrer should predict subsequent forecast revisions, insofar as they do not explain why the forecasts of different forecasters should respond in different ways to the same publicly available information. Our model instead requires that there should be idiosyncratic noise in individual forecasts; for it is only because of the noise term in the law of motion for the memory state (2.19) that the DM is unable to learn the value of μ with perfect precision, and there is no reason for the noise term ω_{t+1}^i to be correlated across decision makers.

It is clear that not only the forecasts of different households, but even the forecasts of professional forecasters exhibit substantial dispersion at a given point in time. A defender

⁴⁹See Figure 1 of Afrouzi *et al.*, who stress this feature of the results of Bordalo *et al.*

of the hypothesis of Bayesian rationality might argue that this simply reflects the fact that different forecasters have access to different private sources of information. Yet experiments in which forecasts are elicited from different subjects who are shown identical sequences of observations indicate that dispersion of forecasts exists even the experimenter can be certain that each subject was exposed to precisely the same information; see in particular Khaw *et al.* (2017) and Afrouzi *et al.* (2020). This indicates noisy cognitive processing of the information presented to the subjects; our model provides at least one possible example of the nature of such idiosyncratic noise in the process by which individuals' expectations are formed.

Our model not only explains why there should exist non-trivial dispersion in the discrepancy between individual forecasts and the consensus forecast, but why this discrepancy should predict subsequent forecast revisions in the way that it does. In our model, all forecasters observe the same external states, but their memories of past states are subject to idiosyncratic noise. In the case of a large enough sample of forecasters, the mean realizations of the memory noise term $\tilde{\omega}_{t+1}^i$ across forecasters i should be close to zero each period, so that (3.7) implies that the mean memory state should be essentially a deterministic function of the past external states,

$$\tilde{m}_{t+1}^{avg} = F_{12} \cdot \sum_{j=0}^{\infty} (\rho_m)^j y_{t-j}.$$

If we let $\tilde{m}_{t+1}^{diff,i} \equiv \tilde{m}_{t+1}^i - \tilde{m}_{t+1}^{avg}$ be the difference between the memory state of DM i and the average memory state, (3.7) implies that

$$\tilde{m}_{t+1}^{diff,i} = \tilde{\omega}_{t+1}^{sum,i}.$$

If we similarly let $\hat{\mu}_t^{diff,i}$ be the difference between DM i 's estimate of μ and the average estimate, it follows from (3.9) that this difference must entirely be due to the difference between i 's memory state and the average memory state, so that we can write

$$\hat{\mu}_t^{diff,i} = \xi \tilde{m}_t^{diff,i}.$$

Finally, it follows from (3.6) that for any forecast horizon h , the difference between i 's forecast and the average forecast will be due entirely to the difference in i 's estimate of μ , so that

$$\hat{y}_{t+h|t}^i - \hat{y}_{t+h|t}^{avg} = (1 - \rho^h) \hat{\mu}_t^{diff,i} = (1 - \rho^h) \xi \tilde{\omega}_t^{sum,i}.$$

In the large-sample limit, the population distribution of realizations of the variable $\tilde{\omega}_{t+1}^{sum,i}$ across forecasters i should be essentially be identical to the distribution of the random variable $\tilde{\omega}_{t+1}^{sum}$. It follows that the distribution of individual forecasts $\{\hat{y}_{t+h|t}^i\}$ should be approximately a Gaussian distribution, with a median very close to its mean, $\hat{y}_{t+h|t}^{avg}$. Thus the consensus forecast should be essentially the same as $\hat{y}_{t+h|t}^{avg}$, and we obtain the prediction

$$\Delta_t^i \equiv \hat{y}_{t+h|t}^i - \hat{y}_{t+h|t}^{cons} = (1 - \rho^h) \xi \tilde{\omega}_t^{sum,i}. \quad (4.6)$$

In a regression of the form (4.5), but where, for purposes of our theoretical derivation, we assume there are no control variables Z_t^i included, the asymptotic value of the coefficient

γ (with a long enough sample) should equal

$$\gamma = \frac{\text{cov}[FR_t^i, \Delta_{t-1}^i]}{\text{var}[\Delta_{t-1}^i]}.$$

The precise predicted value depends on both the degree of persistence of the variable being forecasted and the length of the forecast horizon h , but we can show in general that our model predicts that $-1 < \gamma < 0$.⁵⁰ If we use parameter values $K = 1, \bar{\lambda} = 0.3$, as in Figure 7, the model predicts a regression coefficient on the order of -0.8, as shown in the right panel of Figure 8. And indeed, Fuhrer (2018) finds that the predictable forecast revision over the next quarter should be more than half of the discrepancy between i 's forecast and the consensus forecast (though less than a complete correction of the discrepancy), for the most of the variables and alternative regression specifications that he considers.

5 Conclusion

We have shown that it is possible to characterize the optimal structure of memory, for a class of linear-quadratic-Gaussian forecasting problems, when the cost of a more precise memory is proportional to Shannon's mutual information, and when we assume that the joint distribution of past cognitive states and the memory state is of a multivariate Gaussian form, but with no *a priori* restriction on the dimension of the memory state or the dimensions of past experience that may be more or less precisely recalled. Strikingly, we find that for the class of problems that we consider, the optimal memory structure is necessarily at most one-dimensional. This means that what can be recalled at any time about past observations is simply a noisy recollection of a single summary statistic for past experience. We show how the model parameters determine the law of motion for that summary statistic, and hence what single dimension of past experience will be (imprecisely) available as an input to the DM's forecasts.

Among the implications of our model, two seem of particularly general interest. First, while our formalism allows for the possibility of an independent noisy record of each past observation (as assumed for example in the model of Neligh, 2019), this is not optimal; instead, the optimal memory structure is one in which only a particular weighted average of past observations can be recalled with noise. And second, this weighted average places much larger weights on recent observations than on ones at earlier dates, even though observations at all dates are equally relevant to inference about the value of the parameter μ , which matters for the DM's decisions. Thus our model provides an explanation for "recency bias" in the influence of past observations on current decisions, unlike the model of endogenous memory precision proposed by Afrouzi *et al.* (2020).

We have shown that our model predicts "over-reaction" of forecasts of an autoregressive process to current realizations of the process, of a kind similar to that observed in the forecasts of experimental subjects (Afrouzi *et al.*, 2020), and in survey forecasts of macroeconomic and financial time series (Bordalo *et al.*, 2020a). Moreover, it predicts that the degree of over-reaction should be greater in the case of less persistent time series. This prediction

⁵⁰See Appendix G.4 for an explicit solution for this coefficient.

is confirmed both by laboratory experiments and survey forecasts, as discussed by Afrouzi *et al.* (2020), but cannot be explained by competing explanations for over-reaction, such as the theory of “diagnostic expectations” proposed by Bordalo *et al.* (2018).⁵¹ Our model also predicts that over-reaction should be observed even in the case of time series with a very simple autocorrelation structure — even a white noise time series — unlike the model of “natural expectations” proposed by Fuster *et al.* (2010, 2011), and again in conformity with experimental evidence. And unlike either diagnostic or natural expectations (at least in their most basic forms), our model provides an explanation for heterogeneity in forecasts on the part of forecasters with access to the same information; indeed, the noise in people’s memories is at the heart of our explanation for forecast biases.

In the applications sketched above, we have focused on biases that have been observed in people’s stated expectations. But we suspect that the expectational biases implied by our model can help to explain puzzling aspects of market outcomes as well. For example, Bordalo *et al.* (2020b) argue that a number of well-known puzzles about the behavior of the aggregate stock market are in fact all consistent with a simple dividend discount model of stock prices, under the hypothesis that market expectations regarding firms’ future earnings differ systematically from rational expectations in a particular way, that is furthermore consistent with the biases observed in survey expectations of earnings. They further show that a particular sort of bias in market expectations is needed in order to explain both the biases in survey expectations and the asset pricing anomalies, one very much like the kind of forecast bias predicted by our model.

Briefly, Bordalo *et al.* propose a model in which asset prices at time t are based on market expectations of dividend growth g_{t+h} at various future horizons h . Dividend growth is assumed to be a stationary autoregressive process; market expectations of g_{t+h} differ from rational expectations by an expectational error term $\epsilon_{h,t}$. For any horizon h , $\epsilon_{h,t}$ is assumed to be a stationary, mean-zero autoregressive process, with a substantial degree of persistence; and the innovations in $\epsilon_{h,t}$ are positively correlated with the innovations in g_t , though fluctuations in $\epsilon_{h,t}$ also occur that are uncorrelated with fundamentals. Finally, the fluctuations in $\epsilon_{h,t}$ for different horizons h are perfectly correlated, and $\epsilon_{h,t}$ remains different from zero as $h \rightarrow \infty$, so that innovations in the error process bias expectations about dividend growth in the far future and not only in the near term.

These assumptions are all features of subjective forecasts of the future evolution of the state y_{t+h} in our model (if we identify our y_t with dividend growth). We have shown (in the right panel of Figure 6) that in our model, innovations in y_t cause subjective expectations of the future state to rise more than the RE forecast would, and the effect persists for several periods, though the bias caused by the innovation in any single period t eventually converges to zero. For each horizon h , (3.6) implies that the bias term is equal to $(1 - \rho^h)\hat{\mu}_t$; thus the biases for different forecast horizons are all perfectly correlated. Moreover, as the horizon is increased, the bias term becomes simply $\hat{\mu}_t$ for all large enough h ; thus the forecast errors predicted by the model are above all errors in long-term forecasts.

Our model also implies that there will be random fluctuations in forecast bias that are uncorrelated with any underlying fundamentals; these innovations are indicated by the $\tilde{\omega}_{t+1}$

⁵¹See Afrouzi *et al.* (2020) for a thorough demonstration of the inability of a variety of familiar models to explain this pattern.

shock in (3.5). The most important difference with the reduced-form specification of expectational bias proposed by Bordalo *et al.* is that in their model, there are arbitrary random variations in the “market expectations” that determine the value of the stock market; our model instead implies the existence of idiosyncratic random variation in the beliefs of an individual forecaster, but one might expect that these idiosyncratic variations should cancel out in their effects on the market price. It is possible that a satisfactory model of asset pricing will require us to suppose that some individual traders are large enough for their idiosyncratic beliefs to have a non-negligible effect on aggregate outcomes, as in the model of Gabaix *et al.* (2006). We leave the development of a complete model of asset prices for future work. But it seems likely that imperfect memory of the kind modeled here will be a necessary element in such a model.

References

- [1] Afrouzi, Hassan, Spencer Youngwook Kwon, Augustin Landier, Yueran Ma, and David Thesmar, “Overreaction and Working Memory,” NBER Working Paper no. 27947, October 2020.
- [2] Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, “Diagnostic Expectations and Credit Cycles,” *Journal of Finance* 73: 199-227 (2018).
- [3] Bordalo, Pedro, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer, “Over-Reaction in Macroeconomic Expectations,” *American Economic Review* 110: 2748-2782 (2020a).
- [4] Bordalo, Pedro, Nicola Gennaioli, Rafael LaPorta, and Andrei Shleifer, “Expectations of Fundamentals and Stock Market Puzzles,” NBER Working Paper no. 27283, May 2020b.
- [5] Caplin, Andrew, Mark Dean, and John Leahy, “Rationally Inattentive Behavior: Characterizing and Generalizing Shannon Entropy,” working paper, New York University, February 2019.
- [6] Coibion, Olivier, and Yuriy Gorodnichenko, “Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts,” *American Economic Review* 105: 2644-78 (2015).
- [7] Cover, Thomas M., and Joy A. Thomas, *Elements of Information Theory*, New York: Wiley, 2d ed., 2006.
- [8] Evans, George W., and Seppo Honkapohja, *Learning and Expectations in Macroeconomics*, Princeton: Princeton University Press, 2001.
- [9] Fuhrer, Jeffrey, “Expectations as a Source of Macroeconomic Persistence: Evidence from Survey Expectations in Dynamic Macro Models,” *Journal of Monetary Economics* 86: 22-35 (2017).
- [10] Fuhrer, Jeffrey, “Intrinsic Expectations Persistence: Evidence from Professional and Household Survey Expectations,” Federal Reserve Bank of Boston Working Paper no. 18-9, May 2018.
- [11] Fuster, Andreas, Ben Hébert, and David Laibson, “Natural Expectations, Macroeconomic Dynamics, and Asset Pricing,” *NBER Macroeconomics Annual* 26: 1-48 (2011).
- [12] Fuster, Andreas, David I. Laibson, and Brock Mendel, “Natural Expectations and Macroeconomic Fluctuations,” *Journal of Economic Perspectives* 24(4): 67-84 (2010).
- [13] Gabaix, Xavier, Parameswaran Gopikrishnan, Vasiliki Plerou, and H. Eugene Stanley, “Institutional Investors and Stock-Market Volatility,” *Quarterly Journal of Economics* 121: 461-504 (2006).

- [14] Khaw, Mel Win, Luminita L. Stevens, and Michael Woodford, “Discrete Adjustment to a Changing Environment: Experimental Evidence,” *Journal of Monetary Economics* 91: 88-103 (2017).
- [15] Malmendier, Ulrike, and Stefan Nagel, “Learning from Inflation Experiences,” *Quarterly Journal of Economics* 131: 53–87 (2016).
- [16] Malmendier, Ulrike, Demian Pouzo, and Victoria Vanasco, “A Theory of Experience Effects,” working paper, UC Berkeley, January 2017.
- [17] Metzler, Lloyd A., “The Nature and Stability of Inventory Cycles,” *Review of Economics and Statistics* 23: 113-129 (1941).
- [18] Milani, Fabio, “Expectations, Learning and Macroeconomic Persistence,” *Journal of Monetary Economics* 54: 2065–2082 (2007).
- [19] Milani, Fabio, “Learning and Time-varying Macroeconomic Volatility,” *Journal of Economic Dynamics and Control* 47(C): 94–11 (2014).
- [20] Muth, John F., “Rational Expectations and the Theory of Price Movements,” *Econometrica* 29: 315-335 (1961).
- [21] Neligh, Nathaniel, “Rational Memory with Decay,” working paper, Chapman University, November 2019.
- [22] Sims, Christopher A., “Implications of Rational Inattention,” *Journal of Monetary Economics* 50: 665-690 (2003).
- [23] Slobodyan, Sergey, and Raf Wouters, “Learning in an Estimated Medium-scale DSGE Model,” *Journal of Economic Dynamics and Control* 36: 26–46 (2012).