Optimally Imprecise Memory and Biased Forecasts*

Rava Azeredo da Silveira
ENS and University of Basel

Yeji Sung
Columbia University

Michael Woodford
Columbia University

November 8, 2022

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 the over-reaction to news observed in the laboratory by Afrouzi *et al.* (2020).

*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.,* 2020, and Afrouzi *et al.,* 2020, for recent examples with extensive references to prior literature.)

Here we offer an explanation for the pervasiveness of over-reaction, that depends neither on an assumption that people follow arbitrary (and distinctly sub-optimal) heuristics, or that their forecasts make sense only under an incorrectly specified statistical model. We propose a theory 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 their decision 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.* (2020) and in laboratory forecasting experiments by Afrouzi *et al.* (2020). And unlike the theory of "natural expectations"[1] 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.[1]

In section 1, we present the assumptions of our model of endogenously imprecise memory, and illustrate its consequences for a simple example in which the variable to be forecasted is i.i.d. Section 2 then offers a more general characterization of the optimal memory structure in our model, showing in particular that even when the variable to be forecasted is serially correlated, 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 compares the quantitative predictions of the model to the reported expectations of subjects in the laboratory experiment of Afrouzi *et al.* (2020), 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 that is observed. Section 5 compares our model with alternative explanations for over-reaction of expectations, some of which are based on alternative models of imperfect memory, and section 6 concludes.

1 A Flexible Model of Imprecise Memory: A Simple Example

Here we precisely specify the constraint on the precision of memory that we propose, and illustrate the kind of conclusions that follow from it by first discussing a simple case, in which the state variable to be forecasted is an i.i.d. random variable. The problem of the decision maker [DM] is generalized in the following section.

Suppose that the variable y_t is an independent draw each period from a Gaussian distribution, $y_t \sim N(\mu, \sigma_y^2)$, and that the DM's problem at each time t is to produce a forecast z_t of the value of y_{t+h} — that is, the value of the external state that will be observed h periods later (for some $h \geq 1$). The forecast z_t is produced after observing the value of y_t. If we suppose that the DM's loss from making an inaccurate forecast is proportional to the squared error of the forecast, then an optimal forecasting rule (subject to the memory constraints to be specified below) will be one that minimizes the expected value of the discounted quadratic loss function

$$\sum_{t=0}^{\infty} \beta^t (z_t - y_{t+h})^2, \tag{1.1}$$

[1] Other recent papers that explore the consequences of assuming that memory allows only a noisy recollection of past observations include Afrouzi *et al.* (2020) and Neligh (2022). 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. See section 5.2 for further discussion.

where $0 < \beta < 1$ is the DM's discount factor.

Given that future realizations of the state are completely independent of anything observed in the past, it is obvious that if the parameters of the distribution from which y_t is drawn are known (that is, if the DM's decision rule can be designed using the values of these parameters), then the optimal forecast each period will simply be $z_t = \mu$, the unconditional expected value of y_{t+h}. We assume, however, that the DM's decision rule must be chosen without knowledge of the value of μ; instead, the decision is optimized for a prior over the possible values of μ, $\mu \sim N(0, \Omega)$, for some $\Omega > 0$. In this case, an optimal decision rule will seek to estimate the value of μ (and hence the minimum-mean-squared-error [MMSE] forecast) from observations of the state that have been made up to time t.

We simplify the discussion by supposing that the value of σ_y^2 is known (can be used in specifying the DM's decision rule); this makes it straightforward to say how much can be inferred about the value of μ from an observation of the state y_t. In the case of perfect memory, so that the DM's forecast z_t can be a function of the complete sequence of observations $(y_0, \ldots, y_t)$ from some initial period zero onwards, the computation of the MMSE estimate of μ is a standard Kalman-filtering problem. Posterior beliefs after y_{t-1} has been observed are of the form $\mu \sim N(\hat{\mu}_{t-1}, \hat{\sigma}_{t-1}^2)$, where the mean and variance of this Gaussian distribution are to be calculated. It then follows that after the next observation y_t, the new posterior will be of the same form, with mean and variance given by the recursions

$$\hat{\mu}_t \;=\; \hat{\mu}_{t-1} \,+\, \gamma_t(y_t - \hat{\mu}_{t-1}),$$

$$\hat{\sigma}_t^2 \;=\; \frac{\hat{\sigma}_{t-1}^2 \sigma_y^2}{\hat{\sigma}_{t-1}^2 + \sigma_y^2},$$

where the "Kalman gain"

$$\gamma_t \;=\; \frac{\hat{\sigma}_{t-1}^2}{\hat{\sigma}_{t-1}^2 + \sigma_y^2}$$

is a factor between 0 and 1. These equations can be solved recursively to determine $\hat{\mu}_t$ and $\hat{\sigma}_t^2$ for all t (given a sequence of realizations of the state), starting from initial conditions $\hat{\mu}_{-1} = 0$, $\hat{\sigma}_{-1}^2 = \Omega$.

These equations imply that the precision $\hat{\sigma}_t^{-2}$ grows linearly with the number of observations, and hence that $\hat{\sigma}_t^2 \to 0$ as $t \to \infty$, regardless of the sequence of observations. Eventually the correct value of μ is learned to arbitrary precision, and new observations cease to affect the estimate of μ ($\gamma_t \to 0$), and consequently cease to affect the DM's forecast. Thus forecasts are eventually the same as under an assumption of (full-information) rational expectations. We wish to examine how these conclusions change in the case of imperfect memory.

1.1 Feasible memory structures

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 $(y_0, y_1, \ldots, y_{t-1})$.) We assume that current external state y_t is perfectly observable,[2] 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

$$m_{t+1} \;=\; \Lambda_t s_t \,+\, \omega_{t+1}, \qquad \omega_{t+1} \,\sim\, N(0,\, \Sigma_{\omega,t+1}) \tag{1.2}$$

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 (2022).[3] 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 $(y_t, y_{t-1}, \ldots, y_{t-n})$ 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.[4] In this section, we assume that there is a finite upper bound $\bar{I}$ on the feasible rate of information transmission: thus feasible memory structures must satisfy the constraint $I_t \leq \bar{I}$, where I_t is the mutual information between s_t and m_{t+1}.[5] Subject to

[2] The case in which the current state is observable only imprecisely is discussed in Sung (2022).

[3] Note 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.

[4] Mutual 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).

[5] In section 2, we generalize this assumption to allow I_t to be increased at some positive marginal cost.

this constraint on feasible memory structures, both the memory structure and the decision rule (specifying z_t as a function of the cognitive state s_t) each period are chosen so as to achieve the minimum possible expected value of (1.4).

1.2 The optimal memory structure

Here we sketch the implications of this model of noisy memory for the simple forecasting problem introduced above. (A more complete presentation of the calculations is offered below, where we also discuss a more general problem.) If we introduce the notation

$$\mu \,|s_t \;\sim\; N(\hat{\mu}_t,\, \hat{\sigma}_t^2) \tag{1.3}$$

for the posterior distribution for μ conditional on the cognitive state s_t, we observe that the DM's optimal decision rule will be $z_t = \hat{\mu}_t$ each period, and that the minimum achievable value for the loss function (1.4), given the memory structure, will be

$$\sum_{t=0}^{\infty} \beta^t [\hat{\sigma}_t^2 + \sigma_y^2].$$

It follows that the optimal memory structure will be the one that minimizes the implied value of

$$\sum_{t=0}^{\infty} \beta^t \hat{\sigma}_t^2, \tag{1.4}$$

When $\{y_t\}$ is an i.i.d. random variable, the only possible relevance of memory for decisions in periods $t+1$ or later is the evidence that memory can provide about the value of the parameter μ. Hence the only aspect of the cognitive state s_t that is worth remembering later is what was known then about the value of μ, which is to say, the parameters of the distribution (1.3). Given the linear-Gaussian dynamics in our model, one can show that $\hat{\sigma}_t^2$ is independent of the history of realizations of the external state, and hence the same in all possible cognitive states s_t.[6] Thus the scalar quantity $\hat{\mu}_t$ is the only aspect of the cognitive state that is worth remembering.

Since a memory state m_{t+1} that was informative about any other aspect of s_t would increase the value of I_t without increasing the information provided about the value of μ, an optimal memory structure will make the distribution of the random variable $m_{t+1}\,|s_t$ a function only of $\hat{\mu}_t$. Under the assumption of linear-Gaussian dynamics (1.2), we must therefore be able to write

$$m_{t+1} \;=\; \Lambda_t\, \hat{\mu}_t \;+\; \omega_{t+1}, \tag{1.5}$$

where Λ_t is now a column vector rather than a matrix.

Given the memory state m_t that can retrieved in any period, the implied posterior distribution for the parameter μ will be a Gaussian distribution,

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

[6] It depends on t, which is to say the number of observations that have occurred; but this is assumed to be available as an input to the decision rule, rather than something that has to be remembered using costly memory.

(When memory is imperfect, however, we can no longer identify $(\bar{m}_t,\, \Sigma_t)$ with $(\hat{\mu}_{t-1},\, \hat{\sigma}^2_{t-1})$.) After the value of y_t is observed, these beliefs are updated to a posterior of the form (1.3), where

$$\hat{\mu}_t \;=\; \bar{m}_t \,+\, \gamma_t\,(y_t - \bar{m}_t), \tag{1.6}$$

using the notation

$$\gamma_t \;=\; \frac{\Sigma_t}{\Sigma_t + \sigma_y^2} \tag{1.7}$$

for the Kalman gain, and

$$\hat{\sigma}_t^2 \;=\; \frac{\Sigma_t \sigma_y^2}{\Sigma_t + \sigma_y^2}. \tag{1.8}$$

Our linear-Gaussian framework further implies that $\bar{m}_t$ must be a linear function of m_t, while Σ_t is independent of m_t. It then follows from (1.5) that we can write

$$\bar{m}_{t+1} \;=\; \lambda_t\,\hat{\mu}_t \,+\, \bar{\omega}_{t+1}, \tag{1.9}$$

where now λ_t is a scalar, and $\bar{\omega}_{t+1} \;\sim\; N(0,\, \sigma^2_{\bar{\omega},t+1})$ a scalar random variable. We can further show that for any feasible memory structure, λ_t must be a quantity no less than zero and less than 1, and that

$$\sigma^2_{\bar{\omega},t+1} \;=\; \lambda_t(1-\lambda_t)\,\mathrm{var}[\hat{\mu}_t] \;=\; \lambda_t(1-\lambda_t)\,[\Omega - \hat{\sigma}_t^2].$$

Thus the law of motion (1.9) is fully specified by choosing a value for λ_t.

The only information about s_t contained in m_{t+1} must be the information about the value of $\hat{\mu}_t$ provided by the value of $\bar{m}_{t+1}$; hence under an optimal information structure, the value of I_t will be the mutual information between the random variables $\hat{\mu}_t$ and $\bar{m}_{t+1}$. It follows from (1.9) that this is equal to $-(1/2)\ln(1-\lambda_t)$, an increasing function of λ_t. Thus the constraint $I_t \leq \bar{I}$ can alternatively be expressed as a constraint of the form $\lambda_t \leq \bar{\lambda}$, where $0 < \bar{\lambda} < 1$. (The limiting case in which $\bar{\lambda} = 1$ corresponds to no upper bound on the mutual information, and hence perfect memory.)

We can further show that the uncertainty about the value of μ in all periods $\tau > t$ is minimized (and hence the loss function (1.4) is minimized) by setting λ_t as large as possible, consistent with the constraint. Hence in each period the upper bound constraint will bind, and the optimal memory structure will be the one in which $\lambda_t = \bar{\lambda}$ each period. The law of motion (1.9) can accordingly be written

$$\bar{m}_{t+1} \;=\; \bar{\lambda}\,\hat{\mu}_t \,+\, \bar{\omega}_{t+1}, \tag{1.10}$$

and the associated posterior variance will equal

$$\Sigma_{t+1} \;=\; \Omega - \mathrm{var}(\bar{m}_{t+1}) \;=\; (1-\bar{\lambda})\Omega \,+\, \bar{\lambda}\hat{\sigma}_t^2. \tag{1.11}$$

Equations (1.6)–(1.8) and (1.10)–(1.11) then constitute a complete system of equations to recursively determine the evolution of the variables $\{\bar{m}_t, \Sigma_t, \hat{\mu}_t, \hat{\sigma}_t^2\}$ for all $t \geq 0$ given the sequence of observations $\{y_t\}$, starting from initial conditions $\bar{m}_0 = 0, \Sigma_0 = \Omega$ corresponding to the prior. (This generalizes the recursive system given above for the case of perfect memory.)

1.3 Implications for forecast dynamics and forecast errors

In the simple problem considered here, the optimal forecast each period is given by $z_t = \hat{\mu}_t$; the predictable part of the forecast error (if any) will simply be the predictable error (if any) in $\hat{\mu}$ as an estimate of μ; and the mean squared error of the forecast will equal $\hat{\sigma}_t^2 + \sigma_y^2$, where $\hat{\sigma}_t^2$ is the mean squared error of the estimate of μ. Thus we need only analyze the dynamics of the estimate $\hat{\mu}_t$ and the estimation error that this reflects.

In the perfect-memory case, the recursive system of equations presented above imply that

$$\hat{\mu}_t \;=\; \gamma_t \sum_{\tau=0}^{t} y_\tau,$$

so that each observation up through date t has an equal effect on the estimate (there are no "order effects"), and the optimal estimate is a positive multiple of the mean of the observed values $\{y_\tau\}$. The multiplicative factor $k_t = (t+1)\gamma_t$ is less than 1,[7] but converges to 1 as t becomes large (and $\gamma_t \to 0$).

In the noisy-memory ($\bar{\lambda} < 1$) case, instead, the solution for $\hat{\mu}_t$ is different in three important respects. First, the Kalman gain γ_t (which is again the weight on the current observation y_t) does not converge to 0 as t becomes large, but instead converges to a long-run value $\bar{\gamma}$ between 0 and 1. This is because the dynamics of the sequence $\{\hat{\sigma}_t^2\}$ implied by equations (1.8) and (1.11) imply that $\hat{\sigma}_t^2$ converges to a positive long-run value,[8] so that (1.7) then implies that γ_t converges to a positive value less than 1. Second, the weights on the different observations $\{y_\tau\}$ are not equal; instead the effect of a given observation on the estimate is smaller, the more distant the observation in the past (a "recency effect"). And third, instead of $\hat{\mu}_t$ being a deterministic function of the $\{y_\tau\}$, the estimate is also affected by the sequence of memory noise terms $\{\bar{\omega}_\tau\}$.

Specifically, one can write

$$\hat{\mu}_t \;=\; \sum_{j=0}^{t} \alpha_{j,t} y_{t-j} \;+\; \sum_{j=0}^{t} \beta_{j,t} \bar{\omega}_{t-j}, \tag{1.12}$$

where the coefficients are given by

$$\begin{aligned} \alpha_{j,t} \;&=\; \bar{\lambda}^j \gamma_{t-j} \Pi_{i=1}^{j} (1 - \gamma_{t-j+i}), \\ \beta_{j,t} \;&=\; \bar{\lambda}^j \Pi_{i=0}^{j} (1 - \gamma_{t-j+i}) \end{aligned} \tag{1.13}$$

for all $j \geq 0$.[9] In the limit as t becomes large, the coefficients converge:

$$\alpha_{j,t} \;\to\; \alpha_j \;\equiv\; \bar{\gamma}(\bar{\lambda}(1-\bar{\gamma}))^j,$$

$$\beta_{j,t} \;\to\; \beta_j \;\equiv\; \bar{\lambda}^j (1-\bar{\gamma})^{j+1}.$$

Thus in the large-t limit, the estimate $\hat{\mu}_t$ comes to equal a positive multiple of an exponentially-weighted moving average of past observations $\{y_\tau\}$, plus a serially-correlated

[7] This reflects shrinkage of the Bayesian estimate of μ toward the prior mean of zero.

[8] See Figure 1 below for numerical examples, and Appendix F.4 for further analytical discussion.

[9] When $j = 0$ in (1.13), we define the product with no factors to equal 1.

noise term. Because of the exponentially decreasing weights, the term $\sum_j \alpha_j y_{t-j}$ continues to fluctuate randomly in response to the randomness in recent observations, rather than converging to the true value of μ with probability 1 (as in the perfect-memory case). Because $\sum_j \alpha_j < 1$, this term is also on average closer to 0 than is the true value of μ: the shrinkage toward the prior mean is not eliminated even as $t \to \infty$. And because the β_j are positive, the term $\sum_j \beta_j \bar{\omega}_{t-j}$ is an additional source of random variation in the estimate (and hence in the DM's forecast), independent of the sequence of observations $\{y_\tau\}$.

Because the limiting coefficients α_j are positive, the estimate $\hat{\mu}_t$ (and hence the DM's forecast z_t) continues to be influenced by recent observations y_{t-j} even when t is large — unlike the rational-expectations forecast, $z_t = \mu$. Thus the DM's forecast is predicted to "over-react" to news about recent observations.[10] Though the calculations required are more complex, we obtain qualitatively similar conclusions in the case that the DM forecasts a serially correlated variable, as we show next.

2 The Optimal Memory Structure when the State is Persistent

We now consider a more general class of linear-quadratic decision problems, allowing both for simultaneous forecasting of many different horizons, and for persistent dynamics in the state $\{y_t\}$ that is to be forecasted. We allow the state y_t to follow a stationary AR(1) process. We write its law of motion as

$$y_t \;=\; \mu \,+\, \rho(y_{t-1} - \mu) \,+\, \epsilon_{yt}, \tag{2.1}$$

where μ is again the mean, 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 $N(0,\, \sigma_\epsilon^2)$. 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

$$\mathrm{E}\sum_{t=0}^{\infty} \beta^t (z_t - \tilde{z}_t)' W (z_t - \tilde{z}_t), \tag{2.2}$$

where 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,[11]

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

[10] We compare the predictions of our model to the measures of over-reaction reported by Afrouzi *et al.* (2020) in section 4.

[11] Note 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 $\mathrm{E}_t\tilde{z}_t$, which will generally not be the same as the realized value $\tilde{z}_t$.

where the coefficients $\{A_j\}$ satisfy $\sum_j |A_j| < \infty$. (We again assume that $0 < \beta < 1$.) 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.[12]

To simplify our discussion, we continue to 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 again assumed to be drawn from a 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.

In any problem of this form (regardless of the assumed memory limitations), 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 (2.1) that

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

where we use the notation $\mathrm{E}_t[\cdot]$ for the expected value conditional on the true state at time t, i.e., the value of μ and the history of realizations $(y_0, \ldots, y_t)$, 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

$$z_t \;=\; \sum_{j=0}^{\infty} A_j[\hat{\mu}_t + \rho^j (y_t - \hat{\mu}_t)], \tag{2.3}$$

where we again use the notation (1.3).

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 (2.3), under different possible assumptions

[12] For 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 $j \geq 0$.

about how the estimate $\hat{\mu}_t$ is formed.[13] In the case of any forecasting rule of that kind, the loss function (2.2) is equal to

$$\alpha \cdot \sum_{t=0}^{\infty} \beta^t MSE_t \tag{2.4}$$

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

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

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.

Feasible memory structures are again assumed to be described by linear-Gaussian dynamics of the kind specified in section 1.1. However, rather than assuming that there must be a fixed upper bound $\bar{I}$ on the mutual information I_t, we can assume more generally that there is a cost $c(I_t)$ of storing and/or accessing the memory state m_{t+1}, where $c(I)$ is an increasing and (at least weakly) convex function.[14]

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.[15]

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

$$\sum_{t=0}^{\infty} \beta^t \left[\alpha \cdot MSE_t \,+\, c(I_t),\right] \tag{2.5}$$

[13] See Appendix A for details of the argument.

[14] The case of a fixed upper bound on the mutual information, considered above, can be nested as a special case of this model, in which $c(I) = 0$ for all $I \leq \bar{I}$, while the function is equal to $+\infty$ in the case of any $I > \bar{I}$.

[15] See the discussion in sections 3.4 and 5.2.2. The model of Afrouzi *et al.* also assumes that information that is retrieved from memory (at a cost) for use in a decision at time t has no consequences for the information that will be available at later times; the perfectly accurate record of all past observations continues to contain the same information regardless of what is retrieved at time t, while the information retrieved (added to "working memory") at time t is not available at any later time. This makes the problem of optimal selection of the information to be retrieved at any time t a (relatively simple) static problem in their model, whereas it is a dynamic problem in the model proposed here, since in our model, information not remembered at time t cannot (at any cost) be retrieved in any later period.

taking into account both the cost of less accurate forecasts (2.4) 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}. We turn now to a general characterization of the solution to this dynamic optimization problem.

2.1 Implications of linear-Gaussian dynamics

For any memory structure in the class specified in section 1.1, the posterior distribution over possible values of $(\mu, y_{-1}, y_0, \ldots, y_{t-1})$ 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 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 $\bar{m}_t \equiv \mathrm{E}[x_t\,|m_t]$ 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

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

(Here $\bar{m}_t$ is now a 2-vector, and Σ_t a 2×2 matrix.) And let us furthermore introduce the vectors

$$e_1' \;\equiv\; [1 \quad 0], \qquad c' \;\equiv\; [1-\rho \quad \rho]$$

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 of the full-information forecast $E_{t-1}y_t$.

The for μ after also observing y_t will then be of the form (1.3), with mean and variance given by the usual Kalman filter formulas,[16]

$$\hat{\mu}_t \;\equiv\; \mathrm{E}[\mu\,|s_t] \;=\; e_1'\bar{m}_t \,+\, \gamma_{1t}\,[y_t \,-\, c'\bar{m}_t], \tag{2.6}$$

$$\hat{\sigma}_t^2 \;\equiv\; \mathrm{var}[\mu\,|s_t] \;=\; e_1'\Sigma_t e_1 - \gamma_{1t}^2[c'\Sigma_t c \,+\, \sigma_\epsilon^2], \tag{2.7}$$

with a Kalman gain equal to[17]

$$\gamma_{1t} \;=\; \frac{e_1'\Sigma_t c}{c'\Sigma_t c \,+\, \sigma_\epsilon^2}. \tag{2.8}$$

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

[16] Note that these equations generalize (1.6)–(1.8) above for the $\rho = 0$ case.

[17] We 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.

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

$$\hat{\sigma}_t^2 \;\leq\; \hat{\sigma}_0^2 \;\equiv\; \frac{\Omega\sigma_y^2}{\Omega + \sigma_y^2}, \tag{2.9}$$

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

$$MSE_t \;=\; \hat{\sigma}_t^2. \tag{2.10}$$

This determines the value of one of the terms in (2.5) 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.

2.2 The sufficiency of memory of a reduced cognitive state

We further show in the appendix[18] that an optimal memory structure makes the memory state m_{t+1} a function only of the "reduced cognitive state"

$$\bar{s}_t \;\equiv\; \begin{bmatrix} \hat{\mu}_t \\ y_t \end{bmatrix} \;=\; \mathrm{E}[x_{t+1}\,|s_t]. \tag{2.11}$$

We first note (using (2.6) 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

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

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

$$\bar{m}_{t+1} \;=\; \bar{\Lambda}_t\bar{s}_t \,+\, \bar{\omega}_{t+1}, \tag{2.12}$$

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

[18] See Appendix C for details of the argument.

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 (2.12) to be consistent with the normalization

$$\mathrm{E}[\bar{s}_t \,|\bar{m}_{t+1}] \;=\; \bar{m}_{t+1}. \tag{2.13}$$

We show in the appendix that this relationship will hold if and only if[19]

$$\Sigma_{\bar{\omega},t+1} \;=\; (I - \bar{\Lambda}_t) X_t \bar{\Lambda}_t', \tag{2.14}$$

where $X_t \equiv \mathrm{var}[\bar{s}_t]$. Note that (2.11) implies that

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

from which we see that

$$X_t \;=\; X(\hat{\sigma}_t^2) \;\equiv\; \begin{bmatrix} \Omega - \hat{\sigma}_t^2 & \Omega \\ \Omega & \Omega + \sigma_y^2 \end{bmatrix}. \tag{2.15}$$

Thus the matrix X_t depends only on the value of $\hat{\sigma}_t^2$. In addition, (2.9) 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 (2.14) 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 (2.14) is also symmetric); and (b) the right-hand side of (2.14) 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 (2.14).

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 $w >> 0$ such that $\bar{s}_t = w \cdot y_t$. In this case,

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

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

$$\bar{\Lambda}_t \;=\; \lambda_t \frac{ww'}{w'w}, \tag{2.16}$$

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.[20]

[19] See the introductory section of Appendix D for details of the argument.

[20] Restricting 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 (2.17) below also valid in the case that $X_t = X_0$, and thus it allows us to state certain conditions below more compactly.

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 (2.14). As shown in the appendix,[21] 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 (2.7). 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 (2.9) 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 (2.9), 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 $t \geq 0$ 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 (2.10).

We can also uniquely identify the information cost implied by such a sequence of transition matrices, since as shown in the appendix,[22] the mutual information between s_t and m_{t+1} will be given by

$$I_t \;=\; I(\bar{\Lambda}_t) \;\equiv\; -\frac{1}{2}\log\det(I - \bar{\Lambda}_t) \tag{2.17}$$

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 (2.17) 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, $I - \hat{\Lambda}_t$ 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, $\det(I - \hat{\Lambda}_t)$ 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 (2.5) 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 (2.5). 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.[23]

[21] See Appendix D.1 for details of the argument.

[22] See Appendix D.2 for details of the argument.

[23] Note 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 (2.5) will equal the finite quantity $\hat{\sigma}_0^2/(1-\beta)$.

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

$$\sum_{\tau=t}^{\infty} \beta^{\tau-t} \left[\alpha \cdot MSE_\tau \, + \, c(I_\tau)\right] \tag{2.18}$$

will be well-defined.[24]

We can then consider the problem of choosing an admissible continuation plan $\{\bar{\Lambda}_\tau\}$ for $\tau \geq t$ so as to minimize (2.18), 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 (2.5), given the initial condition for $\hat{\sigma}_0^2$ specified in (2.9).) Let $V(\hat{\sigma}_t^2)$ be the lowest achievable value for (2.18), 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 (2.9), and is independent of the date t from which we consider the continuation problem. Note that the lower bound necessarily lies in the interval

$$\alpha\hat{\sigma}_t^2 \;\leq\; V(\hat{\sigma}_t^2) \;\leq\; \alpha\left[\hat{\sigma}_t^2 \, + \, \frac{\beta}{1-\beta}\hat{\sigma}_0^2\right]. \tag{2.19}$$

(The lower bound follows from the fact that $MSE_t = \hat{\sigma}_t^2$, and all other terms in (2.18) must be non-negative; the upper bound is the value of (2.18) 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

$$V(\hat{\sigma}_t^2) \;=\; \min_{\bar{\Lambda}_t \in \mathcal{L}(X(\hat{\sigma}_t^2))} [\alpha\hat{\sigma}_t^2 \, + \, c(I(\bar{\Lambda}_t)) \, + \, \beta V(f(\hat{\sigma}_t^2, \bar{\Lambda}_t))], \tag{2.20}$$

where $I(\bar{\Lambda}_t)$ is the function defined in (2.17). 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 (2.20). 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 (2.9), and consistent with the bounds (2.19) everywhere on this domain. Then (2.20) defines a mapping $\Phi : \mathcal{F} \to \mathcal{F}$: 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 (2.20) states that the value function that defines the minimum achievable continuation loss must be a fixed point of this mapping: a function such that $V = \Phi(V)$.

[24] Since a finite value for the continuation objective is always possible (see (2.19) below), it is clear that plans that make (2.18) a divergent series cannot be optimal, and can be excluded from consideration.

We can further show that for any function $V \in \mathcal{F}$, the function $\Phi(V)$ defined by the right-hand side of (2.20) is necessarily a monotonically increasing function.[25] 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 (2.20) 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 by (2.9); 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 (2.15). We show in the appendix[26] 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 (2.6). The DM's complete vector of forecasts z_t each period is then given by (2.3).

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 (2.20), 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 predictiing (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 (2.6) 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

$$\hat{\mu}_t \;=\; \zeta_t \;+\; \delta_t' \bar{m}_t. \tag{2.21}$$

[25] See Appendix E.1 for a proof.

[26] See Appendix D.1 for details of the argument.

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 (2.20) 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.[27] Hence the optimal choice for $\bar{\Lambda}_t$ must solve the problem

$$\min_{\bar{\Lambda}_t \in \mathcal{L}(X(\hat{\sigma}_t^2))} I(\bar{\Lambda}_t) \qquad \text{s.t. } f(\hat{\sigma}_t^2, \bar{\Lambda}_t) \, \leq \, \hat{\sigma}_{t+1}^2, \tag{2.22}$$

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 (2.22) is non-empty,[28] the solution $\bar{\Lambda}_t$ to this problem must be at most of rank one. Thus it must be of the special form

$$\bar{\Lambda}_t \;=\; \lambda_t X_t v_t v_t', \tag{2.23}$$

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

$$\tilde{m}_{t+1} \;=\; \lambda_t v_t' \bar{s}_t \,+\, \tilde{\omega}_{t+1}, \qquad \tilde{\omega}_{t+1} \,\sim\, N(0,\, \lambda_t(1-\lambda_t)). \tag{2.24}$$

If $\hat{\sigma}_t^2 = \hat{\sigma}_0^2$, the set $\mathcal{L}(X_0)$ consists only of matrices of the form (2.23), with

$$v_t \;=\; \frac{w}{(\Omega + \sigma_y^2)^{1/2}(w'w)}, \tag{2.25}$$

because of (2.16). 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 (2.22). 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 (2.23), that is also consistent with the constraint in (2.22), but which achieves a lower value of $I(\bar{\Lambda}_t)$.

Let the alternative transition matrix be given by (2.23), with

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

where $\delta_{t+1} \equiv e_1 \,-\, \gamma_{1,t+1}c$ is the vector introduced in (2.21), and let the matrix $\Sigma_{\bar{\omega},t+1}$ be correspondingly modified in the way specified by (2.14). We show in the appendix[29] that

[27] In 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 (2.22) must be among the solutions to the problem on the right-hand side of (2.20), even if it is not a unique solution. In such case, showing that the solution to (2.22) is necessarily a singular matrix suffices to show that we can without any loss impose the further constraint in (2.20) that the matrix $\bar{\Lambda}_t$ must be at most of rank one.

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

[29] See Appendix E.2 for details of the argument.

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 $\hat{\sigma}^2_{t+1}$ will be the same, so that the alternative transition matrix also satisfies the constraint in (2.22).

At the same time, we show in the appendix that the reduction in the complexity of memory cannot increase information costs in any period.[30] The new memory structure consists effectively of a scalar memory state $\tilde{m}_{t+1}$ in each period, which is a multiple of $d'_{t+1}\bar{m}_{t+1}$, 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 $\bar{m}_{t+1}$ 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 $\bar{m}_{t+1}$ were multiples of $d'_{t+1}\bar{m}_{t+1}$; which is to say, only if $\bar{\Lambda}_t$ were already of the special form (2.23).

We conclude, then, that an optimal memory structure must involve a transition matrix in every period of the special form (2.23), 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 (2.24). 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 (2.6) and the value of $\hat{\sigma}^2_{t+1}$, which determines the matrix $X_{t+1} = X(\hat{\sigma}^2_{t+1})$.

For any value $0 \leq \hat{\sigma}^2_t < \hat{\sigma}^2_0$, let $\mathcal{V}(\hat{\sigma}^2_t)$ be the set of vectors v_t satisfying $v'_t X(\hat{\sigma}^2_t) v_t = 1$. In the case that $\hat{\sigma}^2_t = \hat{\sigma}^2_0$, we add the further stipulation that $\mathcal{V}(\hat{\sigma}^2_0)$ consists only of the single vector (2.25). Then given a value for $\hat{\sigma}^2_t$, 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 $v_t \in \mathcal{V}(\hat{\sigma}^2_t)$. These determine a value for $\hat{\sigma}^2_{t+1} = f(\hat{\sigma}^2_t, \lambda_t, v_t)$, where now the function f is defined for any values of its arguments satisfying $0 \leq \hat{\sigma}^2_t \leq \hat{\sigma}^2_0$, $0 \leq \lambda_t \leq 1$, and $v_t \in \mathcal{V}(\hat{\sigma}^2_t)$.

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

$$\bar{f}(\hat{\sigma}^2_t, \lambda_t) \;\equiv\; \min_{v_t \in \mathcal{V}(\hat{\sigma}^2_t)} f(\hat{\sigma}^2_t, \lambda_t, v_t). \tag{2.26}$$

In the appendix,[31] we give an explicit algebraic solution for the optimal v_t for any given values $0 \leq \hat{\sigma}^2_t \leq \sigma^2_0$ and $0 < \lambda_t \leq 1$,[32] and hence for the function $\bar{f}(\hat{\sigma}^2_t, \lambda_t)$. The latter

[30] See Appendix E.2 for details of the argument.

[31] See Appendix E.3 for details.

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

function is also defined when $\lambda_t = 0$, and easily seen to equal $\bar{f}(\hat{\sigma}_t^2, 0) = \hat{\sigma}_0^2$. 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 (2.9).

Moreover, it follows from (2.17) that the mutual information associated with the period t memory structure will be given by

$$I_t \;=\; -\frac{1}{2}\log(1 - \lambda_t), \tag{2.27}$$

just as in the i.i.d. case discussed in section 1. The Bellman equation (2.20) can therefore be written in the simpler form

$$V(\hat{\sigma}_t^2) \;=\; \min_{0 \leq \lambda_t \leq 1} [\alpha\hat{\sigma}_t^2 \;+\; c(-(1/2)\log(1 - \lambda_t)) \;+\; \beta V(\bar{f}(\hat{\sigma}_t^2, \lambda_t))]. \tag{2.28}$$

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 (2.5), 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 (2.5) 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 values of I_t greater than $\bar{I}$ are infeasible, as assumed in section 1. Solution for the optimal memory structure is particularly simple in this case. Because of (2.27), the per-period bound on mutual information can equivalently be written as an upper bound $\lambda_t \leq \bar{\lambda}$, just as in section 1. 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 appendix[33] 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.

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

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

where $\phi(\eta; \bar{\lambda})$ is a function that is independent of the scale factor σ_y.[34]

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 (2.9), 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

$$\frac{1}{\eta_{t+1}} \;=\; \frac{1}{\eta_t} + \frac{1-\rho}{1+\rho}. \tag{3.2}$$

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 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 $K = 1$.[35] 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.

[33] See Appendix F.1 for details of the argument.

[34] See Appendix F.1 for an explicit algebraic solution for this function.

[35] The 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).

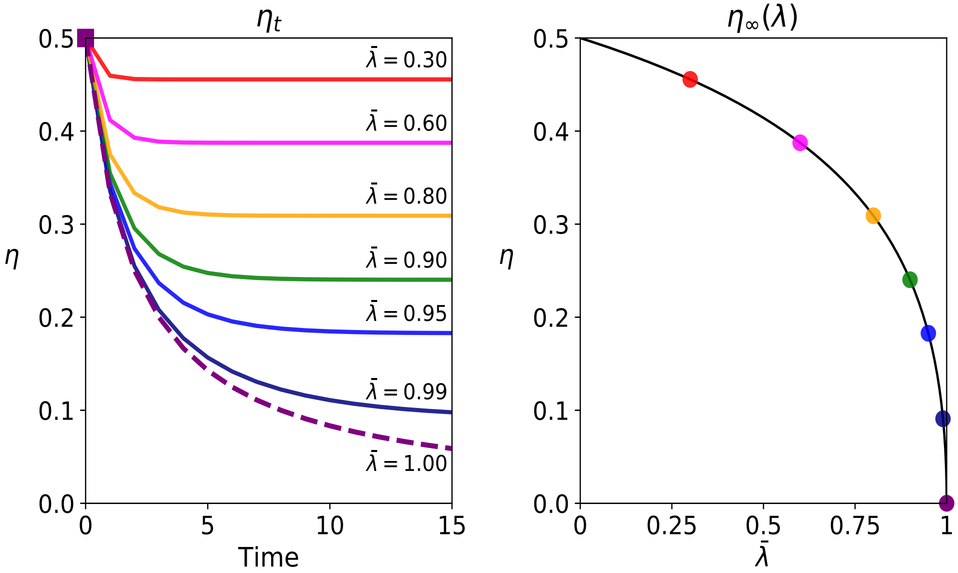


Figure 1: The evolution of scaled uncertainty about μ as the number t of previous (imperfectly remembered) observations grows. The right panel shows the long-run value of scaled uncertainty (to which η_t converges as $t \to \infty$) as a function of the constraint on the complexity of memory, parameterized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

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 (2.26) 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 (2.6), 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

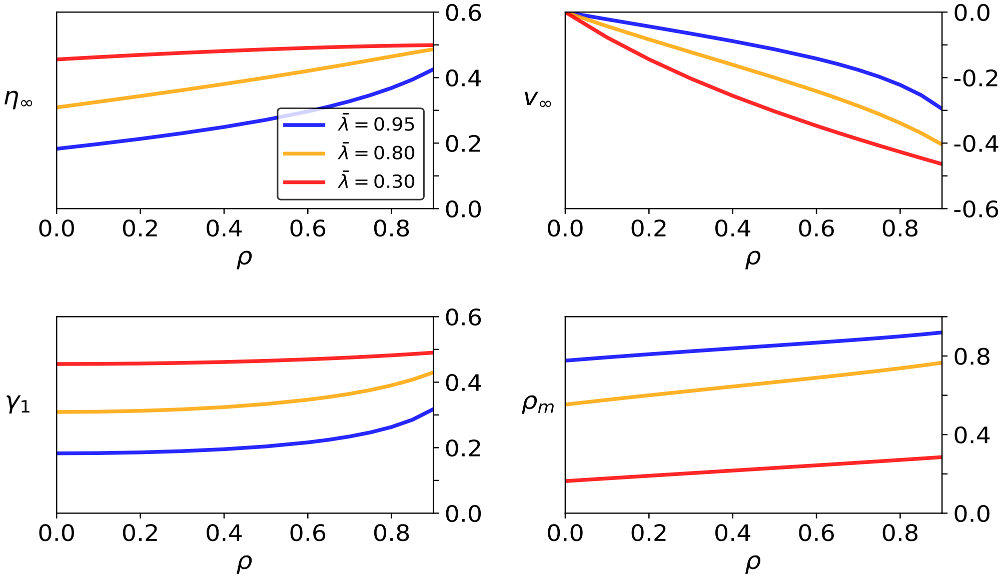


Figure 2: Coefficients describing the optimal memory structure in the long run, as a function of the degree of persistence ρ of the external state, for alternative values of $\bar{\lambda}$. Respective panels show the long-run values for η (measuring uncertainty about μ), the direction vector v (indicating 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).

parameters affects several of the long-run stationary coefficients.[36] 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 components.[37] 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

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

[37] This 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 (2.23)) and hence also of the matrix $\Sigma_{\bar{\omega}}$ (using (2.14)); thus the dynamics of the memory state given by (2.12) are completely specified.

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 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 (2.6). 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 $\bar{\lambda} \to 1$ (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 $\bar{\lambda} \to 0$ (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 $\tilde{m}_\tau$ is proportional to $(\rho_m)^{\tau-t}$ for all $\tau \geq t$;[38] 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

[38] Here we refer to the difference that the realization of $\tilde{\omega}_t$ makes for the forecasts of $\tilde{m}_\tau$ 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.

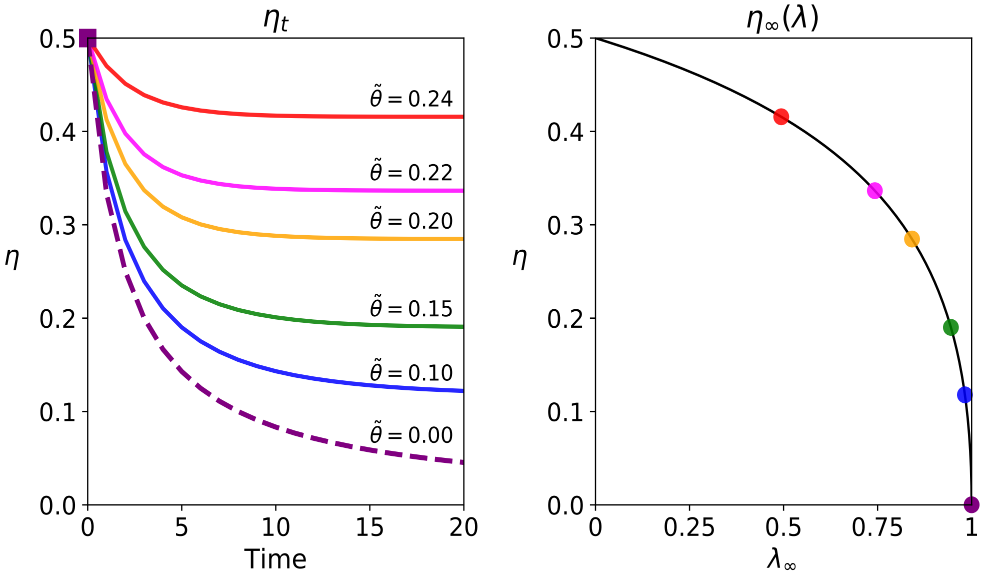


Figure 3: The evolution of scaled uncertainty about μ as the number t of previous (imperfectly remembered) observations grows, now for the case of a linear cost of memory complexity. The right panel shows the long-run value of scaled uncertainty for each value of the cost parameter $\tilde{\theta}$,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

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.

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 $V(\hat{\sigma}^2_{t+1})$ appearing in the Bellman equation (2.28) 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

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

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.

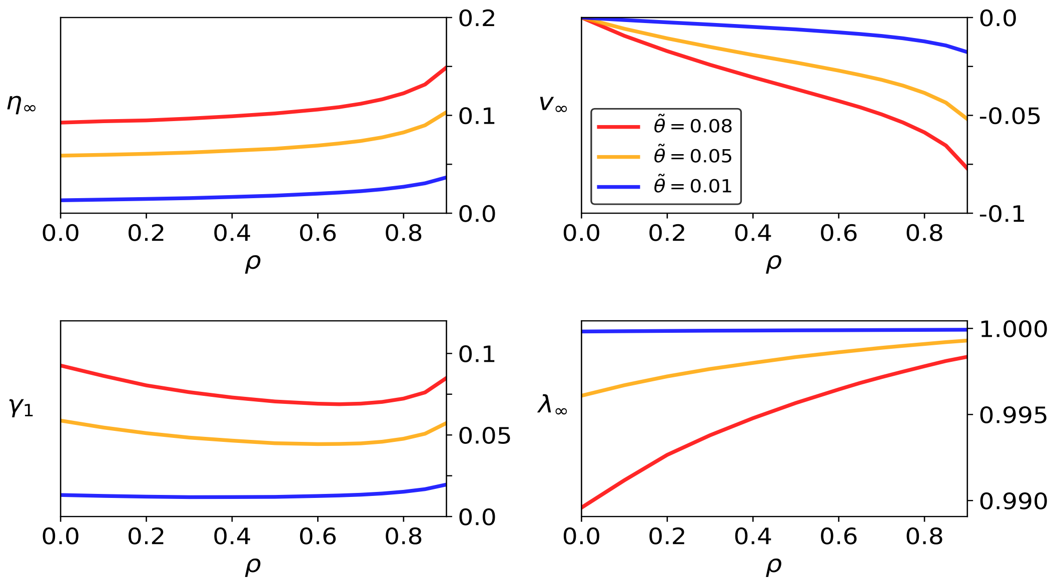


Figure 4: Coefficients describing the optimal memory structure in the long run, as a function of the degree of persistence ρ of the external state, in the case of a linear memory cost function, for alternative values of $\tilde{\theta}$. Respective panels show the long-run values for η, the direction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

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.[39] The policy function $\lambda_t = \lambda^*(\eta_t)$ together with the law of motion

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

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 $\eta_0 = K/(K+1)$.[40] The dynamics of scaled uncertainty as 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

[39] See Appendix F.2 for details.

[40] See Appendix F.2 for further discussion of the implied dynamics.

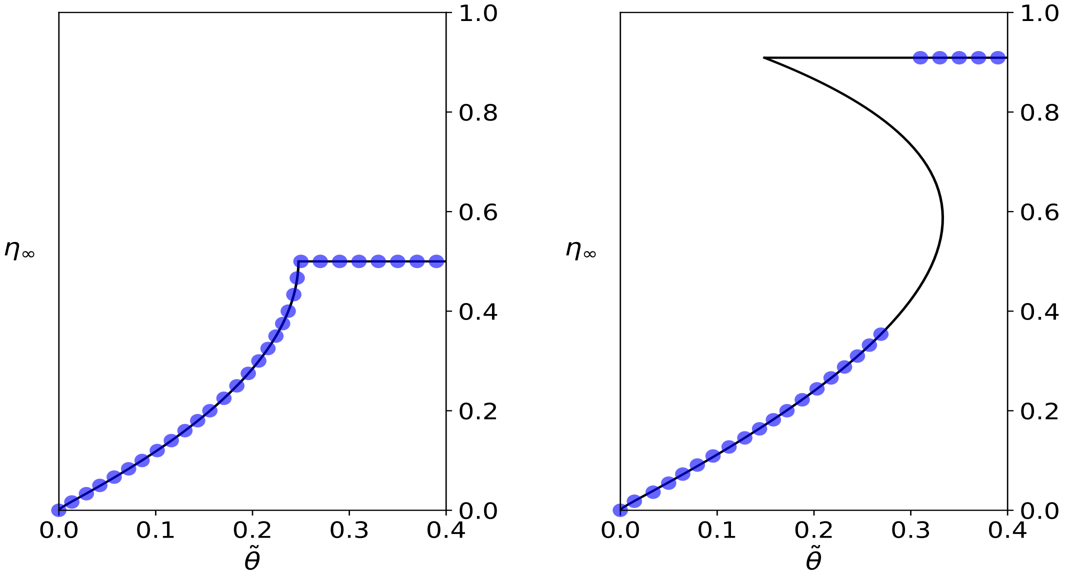


Figure 5: Long-run value of the scaled uncertainty measure η_∞ (blue dots) as a function of the cost parameter $\tilde{\theta}$, in the case of a linear memory cost function. Left panel: $K = 1, \rho = 0$. Right panel: $K = 10, \rho = 0$.

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 $\tilde{\theta}$, 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 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).[41] 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

[41] See Appendix F.4 for derivation of this equation.

result of a small change in parameters;[42] 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 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 $\tilde{\theta}$.[43]

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 (2.1) 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.

[42] See Appendix F.3 for a numerical example.

[43] We 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.

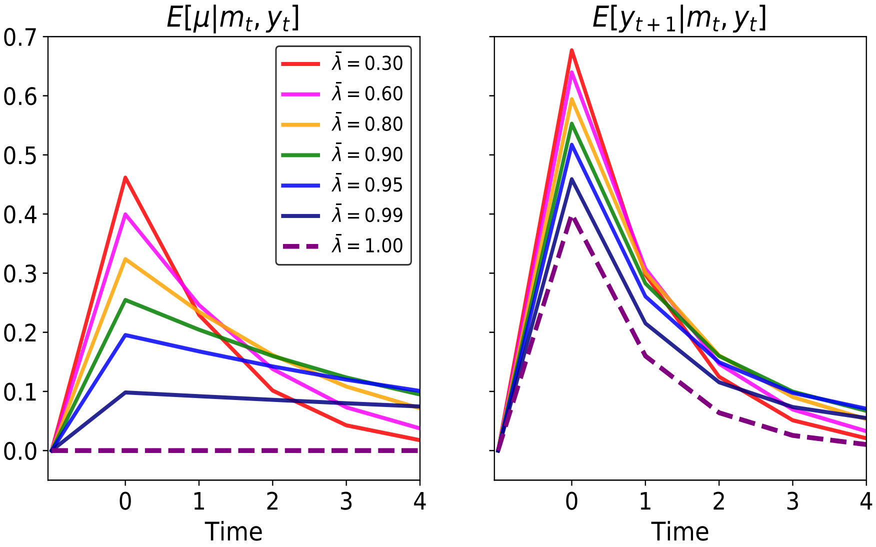


Figure 6: Impulse responses of the DM's estimate of μ (left panel) and one-period-ahead forecast of the state (right panel) to a unit positive innovation in the observed value of y_t at the time marked as "time = 0" on the horizontal axis. Responses are plotted for alternative values of the information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

In the long run, equations (2.1), (2.6), and (2.12) 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

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

where

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

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 (2.6) into (2.24), while the second equation is given by (2.1).

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.[44] 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 $\tau > t$,

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

[44] See Appendix G.1 for the derivation.

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 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 $\rho = 0.4$.[45]

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.[46] 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

[45] See 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.

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

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.

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 and Nagel, 2016; Malmendier *et al.*, 2020) — 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. As we have already previewed in section 1.3, 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 $\tau \leq t$:

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

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

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

is a serially correlated Gaussian noise term.[47]

Equation (2.6) implies that a DM's estimate of the unknown mean μ of the external state is given by a linear relation of the form

$$\hat{\mu}_t \;=\; \xi \tilde{m}_t \;+\; \gamma_1 y_t, \tag{3.9}$$

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

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

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

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 (2.1); 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.[48] The biases in forecasts predicted by our model will 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), as we discuss further in section 5.2.2.

4 Experimental Evidence

We have shown that our model provides an explanation for important qualitative features of observed subjective expectations. Here we briefly discuss the model's quantitative fit with data on subjective expectations from the laboratory experiment of Afrouzi *et al.* (2020). We focus on this particular evidence for a quantitative test of our model, because it involves forecasts of a stationary AR(1) process, and in that sense matches exactly the problem assumed in our theoretical analysis above. A laboratory experiment also has the advantage over field studies of allowing us to be sure exactly what the true data-generating process is, and exactly what information is available to decision makers at each point in time (though of course questions remain about how the situation is understood by the experimental subjects, and what they pay attention to).

As noted in the introduction, Afrouzi *et al.* (2020) conduct a laboratory experiment in which subjects observe successive realizations of an AR(1) process, and forecast what the next realizations should be. They find that subjects' reported 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:

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

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 persistence" measured by the estimated coefficient ρ_h^{subj} and the corresponding coefficient ρ_h

[48] The 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. See further discussion in section 5.1.2 below.

in a regression using actual outcomes:

$$y_{t+h} \;=\; \alpha_h \,+\, \rho_h y_t \,+\, u_{t+h}. \tag{4.2}$$

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 the forecast error $y_{t+h} - \hat{y}_{t+h|t}$ 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 (2.1) 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

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

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

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

(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

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

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

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

We show in the appendix how to calculate this coefficient as a function of the model parameters.[49]

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

[49] See Appendix G.3 for details.

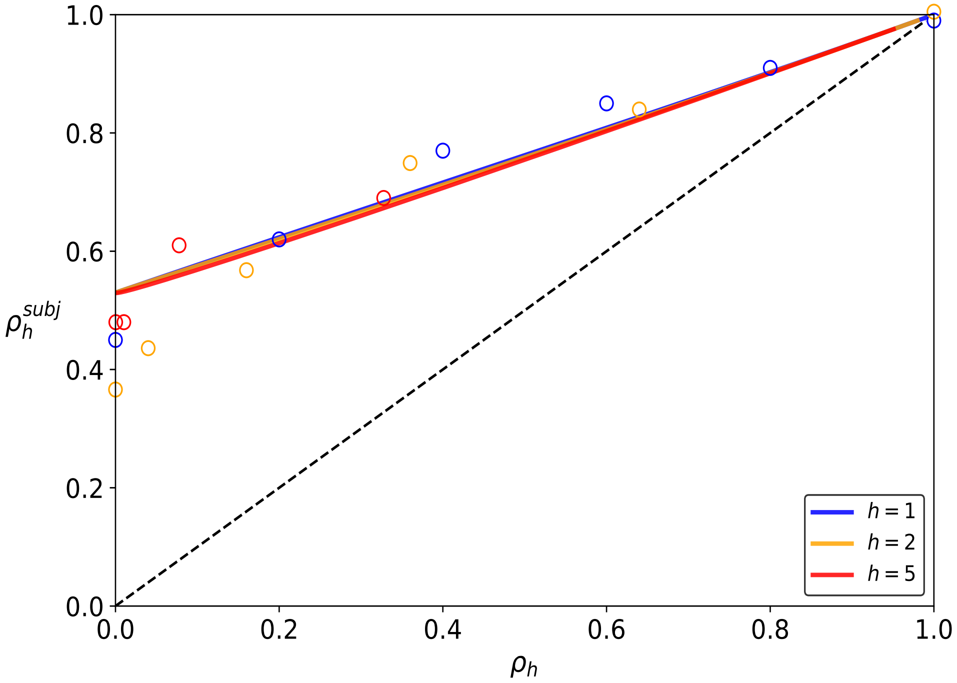


Figure 7: Comparison of the values for the regression coefficients ρ_h and ρ_h^{subj} for different values of ρ and h. (The figure is shown for the case $K = 1, \bar{\lambda} = 0.3$.) The diagonal line indicates the prediction of the rational-expectations hypothesis.

(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 $\rho \to 1$, for any forecast horizon h.[50] 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.[51] 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

[50] This 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 $\tilde{\theta}$ is held constant as ρ is varied (as in Figure 4).

[51] This 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 $\rho \to 1$. (See the plot of their regression coefficients in Figure 7.[52]) 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 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.

There is also evidence of over-reaction to news in the forecasts of macroeconomic and financial variables by professional forecasters, as discussed by Bordalo *et al.* (2020). A satisfactory quantitative account of the predictable forecast errors observed in these forecasts requires an extension of the model presented here, as discussed by Sung (2022). While the more complex model in that paper involves additional information frictions, as addition to allowing for more complex dynamics of the variables that are forecasted, noisy memory of the kind modeled here remains crucial for explaining the observed patterns. And while information frictions of the kind proposed by Coibion and Gorodnichenko (2012, 2015) are also important, Sung finds that quantitative estimates of the size of those frictions are significantly biased by failing to take account of the effects of noisy memory.

5 Related Models

Here we compare our model to alternative models of belief formation that make at least somewhat similar predictions, most notably with regard to the possibility of over-reaction to recent news. We show how our model has important formal similarities to some of these others, and clarify the ways in which it differs from them.

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

5.1 Alternative Explanations for Over-Reaction

We begin by reviewing possible explanations for over-reaction to news that do not rely upon imperfect memory. To simplify the discussion, we here consider only possible explanations for a pattern of over-reaction that would continue to be observed even after an arbitrarily long sequence of observations (rather than discussing transitory dynamics that depend on the DM having insufficient experience with a given context).

5.1.1 Reactions to News when the Mean is Understood to Drift

In our model, over-reaction of forecasts to new observations of the variable y_t reflect revisions of the DM's estimate of the mean of the stochastic process $\{y_t\}$, even though the mean μ is assumed to be constant over time; failure of the DM to learn the exact value of μ, even in the long run, depends on imperfect memory. However, there would be perpetual revision of beliefs about the mean, even with perfect memory (and perfect Bayesian inference) in a world where the mean is (correctly) understood to evolve stochastically over time. In this case, observations farther in the past would be of progressively less relevance to the DM's current estimate of the mean, even with perfect memory.

Such a model can predict forecast dynamics similar (though not identical) to those in our model. As a simple example, suppose that $y_t = \mu_t + \epsilon_t$, where $\epsilon_t \sim N(0,\, \sigma_y^2)$ represents an i.i.d. deviation from the time-varying mean μ_t; and suppose that the mean evolves according to an AR(1) process,

$$\mu_t \;=\; \phi\mu_{t-1} \,+\, \nu_t,$$

where $0 < \phi < 1$ and $\nu_t \sim N(0,\, (1-\phi^2)\Omega)$ is another i.i.d. process. Note that this specification implies that the unconditional prior distribution for the mean is given by $N(0,\, \Omega)$, just as in our model.[53] Let us consider the evolution of the beliefs of a perfect Bayesian DM in such an environment, who observes y_t with perfect precision each period, starting from prior beliefs about μ_0 (before y_0 is observed) corresponding to the unconditional prior.

The posterior distribution for the value of μ_t, just before the observation of y_t, will be a Gaussian distribution $N(\hat{\mu}_{t|t-1},\, \hat{\sigma}^2_{t|t-1})$. The posterior after observing y_t will be another Gaussian distribution $N(\hat{\mu}_{t|t},\, \hat{\sigma}^2_{t|t})$. The mean and variance of this distribution are given by the same Kalman-filter formulas (1.6)–(1.8) as above.[54] In the case of perfect memory, these posterior beliefs about μ_t imply a posterior distribution for the value of μ_{t+1} that is also Gaussian, with mean and variance

$$\hat{\mu}_{t+1|t} \;=\; \phi\hat{\mu}_{t|t}, \tag{5.1}$$

$$\hat{\sigma}^2_{t+1|t} \;=\; \hat{\sigma}^2_{t|t} \,+\, (1-\phi^2)(\Omega - \hat{\sigma}^2_{t|t}). \tag{5.2}$$

Comparison of equation (5.2) with the corresponding equation (1.11) for the dynamics of posterior uncertainty in our noisy-memory model reveals that the degree of uncertainty,

[53] The difference is that in the drifting-mean model, we no longer assume that a value of the mean is drawn from this distribution and then remains constant forever after. Our specification in the previous sections can be regarded as the $\phi \to 1$ limit of this prior.

[54] Here we must substitute $\hat{\mu}_{t|t-1}$ for the prior mean $\bar{m}_t$ in equation (1.6), and $\hat{\sigma}^2_{t|t-1}$ for the prior variance Σ^μ_t. Similarly $\hat{\mu}_{t|t}$ and $\hat{\sigma}^2_{t|t}$ correspond to the variables called simply $\hat{\mu}_t$ and $\hat{\sigma}^2_t$ in the previous equations.

after any particular number of observations of y_t, is the same in both models in the case that $\phi^2 = \bar{\lambda}$. The same path for uncertainty about the mean then implies the same path for the Kalman gain γ_t, given by (1.7). Hence the perfect-Bayesian model with a stochastic mean is equally capable of explaining why a DM's estimate of the mean should continue to be influenced by recent observations, even after a long sequence of observations; the predictions of the two models about this are identical, if the parameter ϕ is chosen appropriately.

However, this does not mean that the two models are observationally equivalent. Equations (5.1)–(5.2) together with the Kalman-filter equation (1.6) imply that after any finite sequence of observations the Bayesian estimate of the mean will be given by a solution of the form (1.12), but with the weights $\{\alpha_j\}$ given by

$$\alpha_{j,t} \;=\; \phi^j \gamma_{t-j} \Pi_{i=1}^{j} (1 - \gamma_{t-j+i})$$

and the weights $\beta_{j,t} = 0$. If we assume that $\phi^2 = \bar{\lambda}$, so that the Kalman gains implied by both models are the same, the weights on past observations $\{y_\tau\}$ will equal

$$\alpha_{j,t} \;=\; \bar{\lambda}^{j/2} \gamma_{t-j} \Pi_{i=1}^{j} (1 - \gamma_{t-j+i}).$$

The weights decay exponentially, like the weights (1.13) implied by the noisy-memory model; but they do not decay at the same rate. (The weights decay more rapidly as j increases in the case of the noisy-memory model; hence the weights are smaller for all $j \geq 1$ in that case.)

Another difference between the two models is that the perfect-Bayesian model implies that there should be a tight relationship between the degree of persistence of the series $\{\mu_t\}$ — and hence the autocorrelation of the observed series $\{y_t\}$ — and the coefficients (such as the Kalman gain γ_t) that describe the dynamics of beliefs about the mean. In the noisy-memory model, the coefficient $\bar{\lambda}$ that determines the size of the Kalman gain and the intrinsic persistence of the belief state can be specified independently of the time-series properties of the process $\{y_t\}$. This flexibility is important for accounting for observed beliefs. Bayesian models of subjective forecasts often have to posit a DM with an apparent prior belief that an unknown state fluctuates more than is actually the case.[55] The noisy-memory model can account for such findings without having to suppose that people fail to learn the correct statistics of their environment. The example just presented shows that noisy memory ($\bar{\lambda} < 1$) can result in belief dynamics similar to those of a Bayesian model in which the DM's prior assumes that the mean μ_t is less persistent than it really is (the prior assumes that $\phi < 1$ when actually μ never changes).

Finally, the perfect-Bayesian model implies that the DM's estimate $\hat{\mu}_{t|t}$ at any time (and hence their forecasts) will be a deterministic function of the sequence of values $(y_0, \ldots, y_t)$ that have been observed. It follows from this that all forecasters who observe the same series should have identical forecasts, and that the variation over time in their forecasts can be fully accounted for by the variation in the values that have been observed. The noisy-memory model instead implies that each DM's beliefs (and hence their forecasts) are affected by memory noise ($\beta_{j,t} \neq 0$); this implies both that forecasts are not perfectly predictable from the past history of the series being forecasted, and that they should differ across forecasters.

[55] See, for example, Yu and Cohen (2009).

This is an attractive feature of the noisy-memory model, since observed forecasts have both of these properties.[56]

5.1.2 Constant-Gain Learning

One also obtains a prediction of perpetual learning, and hence continued over-reaction to news even after an arbitrarily long sequence of observations, in a model where the DM is assumed to estimate the value of the parameter μ using a "constant-gain" variant of least-squares learning (Evans and Honkapohja, 2001, sec. 7.4). Constant-gain (CG) algorithms effectively put an exponentially decreasing weight on observations farther in the past; for example, an unknown mean is estimated by a linear estimator of the form

$$\hat{\mu}_t \;=\; \sum_{j=0}^{t} \gamma(1-\gamma)^j y_{t-j} \;+\; (1-\gamma)^{t+1}\hat{\mu}_{-1},$$

where $0 < \gamma < 1$ is the constant "gain factor" and $\hat{\mu}_{-1}$ is an initial condition (representing the state of belief before y_0 is observed). If we set $\hat{\mu}_{-1} = 0$ (in accordance with the prior assumed in our noisy-memory model), this is similar to the kind of estimate of the unknown mean implied by the perfect-Bayesian model in the case of a drifting mean.[57] Moreover, in the CG algorithm, the value of γ can be specified independently of the dynamics of the process $\{y_t\}$ that is forecasted.[58]

To the extent that a model of CG learning is considered to be empirically realistic, however, a question arises as to what determines the value of the gain parameter. In the adaptive control literature, such algorithms are proposed as a way of dealing with drift in the values of parameters to be estimated; the appropriate value of the gain parameter should thus depend on one's prior regarding the degree of volatility of the parameters to be estimated, as in our discussion above of Bayesian inference when the mean drifts. But once again, the gain parameters that are found to best fit expectations data do not seem to correspond to ones that would be optimal given the degree of structural change in the forecasted time series.[59] Alternatively, authors such as Malmendier and Nagel (2016) propose that aggregate dynamics similar to those predicted by a model of CG learning can result from aggregation of the decisions of people of different ages, who each form beliefs on the basis of their personal experience (and hence on the basis of samples extending different distances into the past).[60] But also under this explanation for CG learning, the predicted

[56] See Sung (2022) for discussion of the difference between individual professional forecasters' forecasts and the consensus forecast, in the case of a variety of macroeconomic variables.

[57] Note that if we consider a limiting case in which $\phi \to 1$ while $(1-\phi^2)\Omega \to \sigma_\nu^2 > 0$, then as $t \to \infty$ the solution (1.12) for the perfect-Bayesian model approaches one in which $\gamma_{t-j} \to \bar{\gamma}$, a constant value between 0 and 1, for all j. In this case the dynamics of the mean estimate implied by the perfect-Bayesian model are exactly those of a constant-gain mean estimate with a gain factor of $\bar{\gamma}$.

[58] In empirical applications (e.g., Milani, 2007, 2014; Slobodyan and Wouters, 2012), the gain parameter and the parameters specifying the persistence of the exogenous states are treated as independent free parameters to be estimated.

[59] See, e.g., Branch and Evans (2006) and Berardi and Galimberti (2017).

[60] For additional examples, see Nakov and Nuño (2015), Schraeder (2016), Collin-Dufresne *et al.* (2017), Ehling *et al.* (2018), and Malmendier *et al.* (2020).

gain parameter should depend on other features of the model, that may not justify a gain parameter as large as the one required to explain the observed degree of over-reaction to news.[61] Our model provides an alternative foundation for belief dynamics similar to those implied by a CG algorithm, in which a substantial gain parameter can exist even when the value of the mean remains constant (or nearly constant) over long periods of time, and even when forecasts have long personal histories of observations.

5.1.3 Forecasts Based on an Incorrect Model

A longstanding explanation for systematic over-reaction to news is the hypothesis that people form their forecasts on the basis of an incorrect statistical model — for example, under an assumption that the fluctuations in $\{y_t\}$ are more persistent than is actually the case. Explanations of this kind have continued to be prominent in the recent literature (e.g., Angeletos *et al.,* 2021), but they raise the question: why should people persist in mis-estimating the dynamics?

Fuster *et al.* (2010, 2011) offer one answer: people's forecasts are optimal, given their estimated model of the dynamics, and their estimated model is the one that best fits the autocorrelation function of the actual series, within some parameteric family of possible models (that need not include the true data-generating process). Their hypothesis of "natural expectations" assumes that the class of statistical models considered is that of all possible AR(k) models, for some fixed bound on k.[62] 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 possible explanatory variables correspond only 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 as discussed above, 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 when the process to be forecasted is white noise.

Like the hypothesis of "natural expectations," our model assumes that forecasts are

[61] Thus Malmendier *et al.* (2020) posit an exponentially decaying influence of earlier experiences on a given DM's expectations, even among the events that have occurred during their lifetime, rather than relying upon demographics alone to account for the rate at which past events cease to influence current market pricing.

[62] More general versions of this hypothesis are considered in the more recent work of Molavi (2022).

optimal, among those forecasting rules in which the forecast is based on only a limited summary of past history; but the way in which we model the limit on the complexity of possible representations of past data is different. Our approach does not impose any *a priori* restriction on either the dimensionality of the memory state or the number of past observations that can be (imperfectly) represented by the memory state. And the form of complexity limit that we assume has the advantage of implying forecasting bias (and more specifically over-reaction) even when the true dynamics are very simple — indeed, even when the true dynamics are white noise (and are recognized by the DM to be white noise).

5.2 Alternative Models of Imprecise Memory

We are also not the only authors to have proposed that expectational biases may result from forecasts being based on imperfect memory of past observations. Here we briefly discuss similarities and differences of alternative proposals with our own approach.

5.2.1 Models of Quasi-Bayesian Belief Updating

Nagel and Xu (2022) propose that a variety of asset-pricing anomalies can be explained by the biases in expectations regarding future asset returns implied by a particular type of departure from perfect Bayesian inference from observed past returns, which they call a model of "fading memory." As in this paper, they consider a situation in which a DM (an investor) must infer the mean μ of a process $\{y_t\}$, based on past observations of this process; and (as in the simple case analyzed in section 1) they assume that the process is i.i.d. (and known to be), and that the only unknown parameter of the distribution is μ. Given a prior $p(\mu)$ over possible values of μ, and a likelihood $p(y\,|\mu)$ for the observation of y_t in any period conditional on the unknown mean, the Bayesian posterior distribution conditional on a finite sequence of observations $\mathbf{y} = (y_{t_0}, \ldots, y_{t-2}, y_{t-1}, y_t)$ is given by

$$p(\mu\,|\boldsymbol{y}) \;\sim\; p(\mu) \prod_{j=0}^{t-t_0} p(y_{t-j}\,|\mu).$$

The Nagel-Xu model of "fading memory" instead assumes a subjective posterior of the form

$$p(\mu\,|\boldsymbol{y}) \;\sim\; p(\mu) \prod_{j=0}^{t-t_0} p(y_{t-j}\,|\mu)^{(1-\nu)^j}, \tag{5.3}$$

for some small quantity $\nu > 0$, which indicates the rate at which memory of past observations "fades." (Note that their model reduces to perfect Bayesian inference in the limiting case in which $\nu = 0$.)

The Nagel-Xu model, like ours, is one in which there is perpetual learning: in the limit as $t_0 \to -\infty$, the posterior distribution (5.3) remains non-degenerate, despite being based on a sample of infinite length. As in our case, the reason is that past observations have a progressively weaker influence on the posterior, the farther they are in the past, and more specifically the influence decreases as an exponential function of the elapsed time. Also as in our case, the Nagel-Xu model implies that one should observe "recency effects." Another important similarity between their approach and ours is that Nagel and Xu model the DM's complete posterior at each point in time, not just the DM's point estimate of μ; and like us, they tie the rate of decay of past information to cognitive limitations, rather than the rate at which the environment is objectively likely to have changed.

Our model differs from that of Nagel and Xu, however, in offering an explicit representation of the imprecise information contained in memory, and then deriving the DM's subjective posterior from (correct) Bayesian conditioning on this imprecise record, rather than directly assuming a particular modification of the Bayesian expression for the posterior beliefs. This is not simply a matter of having failed to provide intermediate steps in the derivation; the subjective beliefs assumed by Nagel and Xu are not correct conditional beliefs, if one were to condition on the information about past observations reflected in the assumed beliefs (and therefore revealed by the DM's cognitive state, since the subjective posterior must be some function of the cognitive state).[63] Our model also differs from theirs in that it implies that individuals' beliefs involve idiosyncratic cognitive noise; thus our model, unlike that of Nagel and Xu, predicts that investors should have heterogeneous beliefs even if they observe identical information. (This difference is relevant for applications to financial economics, since our model of heterogeneous beliefs on the part of individual investors provides a motive for trading, even when all information about asset fundamentals is public.) In these respects, the predictions of our model are not quantitatively identical to those of the model of Nagel and Xu, despite many similarities.

Prat-Carrabin *et al.* (2021) derive a quasi-Bayesian posterior very similar to the one postulated by Nagel and Xu from a hypothesis of "costly Bayesian inference," in which belief updating after each new piece of evidence arrives is distorted (relative to exact Bayesian updating) so as to reduce the precision of the resulting belief state.[64] This hypothesis is even more closely related to the one that we propose here, insofar as the sensitivity of beliefs to past observations decreases over time as a consequence of a cost of storing a more precise record of the DM's past cognitive state. The model of Prat-Carrabin *et al.* differs from ours in identifying the imprecise memory state with the DM's (distorted) posterior beliefs given the sequence of observations to that point; instead, we distinguish between the DM's cognitive state (which includes the memory state m_t) and the probability beliefs that would optimally be inferred from such a state. Thus again, while there are many similarities between the predictions of their model and ours, the predictions are not identical.

5.2.2 Alternative Models of Noisy Memory

Neligh (2022) proposes a model of decaying memory that is conceptually closer to our own in that, as in this paper, it is assumed that memory can be retrieved only with noise, and the judgments that are made are optimal (consistent with correct Bayesian inference) subject to being based on the noisy memory state. The difference with our model is in the way that the memory state, and the cost of retrieving a more precise memory, are modeled. As noted above, Neligh assumes an "episodic" memory, in which there is an independent noisy record of each of the past observations y_τ for $0 \leq \tau \leq t-1$; the element of the memory vector corresponding to the observation at time τ is equal to the value of y_τ plus a mean-zero

[63] See the Appendix, section XX, for detailed discussion.

[64] Prat-Carrabin *et al.* (2022) fits the model to an experimental data set.

Gaussian noise term, distributed independently of the value of y_τ, and with a variance that depends on the amount of elapsed time. This is a special case of the kind of noisy memory that our framework allows for, but is not the form of memory that is found to be optimal for the decision problem considered in this paper. In addition to imposing the constraint that memory must take this form, Neligh endogenizes the precision of memory in only one respect: the precision with which observation y_τ is initially encoded at date τ is optimized (subject to a cost of greater encoding precision), but given the choice of an initial encoding precision, the precision of the memory that can be retrieved after a time delay is exogenously determined by the amount of time that has elapsed. Our model instead allows the precision with which memory is maintained over time to be endogenously varied.[65]

An important similarity between Neligh's model and ours is that in both models, observations more distant in the past are retrieved with greater noise, because of the way in which noise is cumulatively added as the memory state is maintained over time. This means that both models predict recency effects; and it would be possible to specify the rate of increase of memory noise with the passage of time in Neligh's model in such a way as to make the distribution of $\mathrm{E}[\mu \,|m_t]$ conditional on the sequence of past observations — and hence the conditional distribution of all of the DM's forecasts, in the decision problem considered here — the same as the one predicted by our model. There would remain, however, two important differences between Neligh's model and ours. One is that our model derives its predictions from less special assumptions and involves fewer free parameters; thus in the case that both models were equally consistent with empirical observations like those of Afrouzi *et al.* (2020), our model would provide a more parsimonious explanation. And second, Neligh's model implies a much higher-dimensional memory state than does ours. In the case of the decision problem considered in this paper, this makes no difference, as forecasts depend on memory only through a single scalar summary statistic; but the predictions of the two models would likely be different in the case of more complex decisions.

Like us, Afrouzi *et al.* (2020) propose to explain the biases in their experimental subjects' forecasts using a model of endogenously imprecise memory. However, in their model, all past observations are stored in memory with perfect precision; imprecision enters only when an imperfect representation of the contents of memory is retrieved in order to inform a decision. The nature of the imprecise representation that is used for the decision is optimized subject to a cost of precision, which as in our model is based on mutual information (for them, the mutual information between the complete contents of memory and the imprecise representation). As in our model, the information cost implies that an accurate estimate of the value of μ cannot be made on the basis of memory, even after a very large number of observations. Hence subjects' forecasts (assumed as in our model to be optimal subject to having to be conditioned on an imprecise cognitive state) are based on a precise observation of the current y_t together with an imprecise estimate of μ deriving from an imprecise summary of past observations. This results, as in our model, in a prediction of over-reaction to the most recent observation (that can be observed with greater precision than any past observations are recalled); and the predicted degree of over-reaction is greatest in the case of variables with low persistence (since in this case optimal forecasts are largely determined by the optimal

[65] In addition to considering a different class of possible memory structures, Neligh (2022) addresses largely distinct questions from those analyzed here.

estimate of μ).

Despite these similarities in the predictions of the two models, there is an important difference between the model of noisy memory in Afrouzi *et al.* (2020) and our own. Their model implies that all past observations are accessible with equal precision when a forecast needs to be made; hence the optimal noisy representation of the past weights past observations to the extent that they are relevant to the current decision, which implies much less "decay" of old observations than in our model. As a simple example, consider the case in which y_t is i.i.d. Then the contents of memory will be distributed independently of the current observation y_t, and the equally-weighted sample mean of the observations $\{y_\tau\}$ for $0 \leq \tau < t$ will be a sufficient statistic for the information about the mean μ that is contained in the previous observations; hence the optimal representation will be a noisy read-out of this sample mean. It follows that any past observation y_τ (for $\tau < t$) should have exactly the same effect on forecasts at time t as any other: there will be no "recency effect" at all, except for the fact that the observation y_t will have a larger effect than any of the observations at dates $\tau < t$. Thus the model of Afrouzi *et al.* provides no explanation for the kind of recency effects that have frequently been documented in the experimental literature (e.g., Hogarth and Einhorn, 1992), as well as in macroeconomic and financial contexts by authors such as Malmendier *et al.* (2020).

6 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 proptional 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, 2022), 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, and that the degree of over-reaction should be greater in the case of less persistent time series, as observed in the forecasts of experimental subjects (Afrouzi *et al.,* 2020). The same mechanism provides a potential explanation for the frequent observation of over-reaction to news in survey forecasts of macroeconomic and financial time series (e.g., Bordalo *et al.,* 2020). Sung (2022) extends our model to allow for imprecise awareness of the current external state y_t, in addition to the imprecise awareness of the DM's own past cognitive states modeled in this paper, and shows that with this extension the model can account quantitatively for the predictable errors in professional forecasts of a variety of macro variables. In particular, she shows that the model can simultaneously account for the apparent "under-reaction" of consensus forecasts stressed by Coibion and Gorodnichenko (2012, 2015) and the apparent "over-reaction" stressed by Bordalo *et al.* (2020).

In these applications we have focused on biases 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.* (2022) 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 expectional 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 $h \to \infty$, 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 forescast 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 $\tilde{\omega}_{t+1}$ 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] Angeletos, Marios, Zhen Huo, and Karthik Sastry, "Imperfect Macroeconomic Expectations: Evidence and Theory," *NBER Macroeconomics Annual* 35: 1-86 (2021).

[3] Berardi, Michele, and Jaqueson K. Galimberti, "Empirical Calibration of Adaptive Learning," *Journal of Economic Behavior and Organization* 144: 219-237 (2017).

[4] Bordalo, Pedro, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer, "Over-Reaction in Macreconomic Expectations," *American Economic Review* 110: 2748-2782 (2020).

[5] Bordalo, Pedro, Nicola Gennaioli, Rafael LaPorta, and Andrei Shleifer, "Belief Over-Reaction and Stock Market Puzzles," working paper, Oxford Said Business School, April 2022.

[6] Branch, William A., and George W. Evans, "A Simple Recursive Forecasting Model," *Economics Letters* 91: 158-166 (2006).

[7] Caplin, Andrew, Mark Dean, and John Leahy, "Rationally Inattentive Behavior: Characterizing and Generalizing Shannon Entropy," working paper, New York University, February 2019.

[8] Coibion, Olivier, and Yuriy Gorodnichenko, "What Can Survey Forecasts Tell Us About Information Rigidities?" *Journal of Political Economy* 120: 116-159 (2012).

[9] 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).

[10] Collin-Dufresne, Pierre, Michael Johannes, and Lars A. Lochstoer, "Asset Pricing when 'This Time Is Different'," *Review of Financial Studies* 30: 505-535 (2017).

[11] Cover, Thomas M., and Joy A. Thomas, *Elements of Information Theory,* New York: Wiley, 2d ed., 2006.

[12] Ehling, Paul, Alessandro Graniero, and Christian Heyerdahl-Larsen, "Asset Prices and Portfolio Choice with Learning from Experience," *Review of Economic Studies* 85: 1752-1780 (2018).

[13] Evans, George W., and Seppo Honkapohja, *Learning and Expectations in Macroeconomics,* Princeton: Princeton University Press, 2001.

[14] Fuster, Andreas, Ben Hébert, and David Laibson, "Natural Expectations, Macroeconomic Dynamics, and Asset Pricing," *NBER Macroeconomics Annual* 26: 1-48 (2011).

[15] Fuster, Andreas, David I. Laibson, and Brock Mendel, "Natural Expectations and Macroeconomic Fluctuations," *Journal of Economic Perspectives* 24(4): 67-84 (2010).

[16] Gabaix, Xavier, Parameswaran Gopikrishnan, Vasiliki Plerou, and H. Eugene Stanley, "Institutional Investors and Stock-Market Volatility," *Quarterly Journal of Economics* 121: 461-504 (2006).

[17] Hogarth, Robin, and Hillel Einhorn, "Order Effects in Belief Updating: The Belief-Adjustment Model," *Cognitive Psychology* 24: 1-55 (1992).

[18] Malmendier, Ulrike, and Stefan Nagel, "Learning from Inflation Experiences," *Quarterly Journal of Economics* 131: 53–87 (2016).

[19] Malmendier, Ulrike, Demian Pouzo, and Victoria Vanasco, "Investor Experiences and Financial Market Dynamics," *Journal of Financial Economics* 136: 597-622 (2020).

[20] Milani, Fabio, "Expectations, Learning and Macroeconomic Persistence," *Journal of Monetary Economics* 54: 2065–2082 (2007).

[21] Milani, Fabio, "Learning and Time-varying Macroeconomic Volatility," *Journal of Economic Dynamics and Control* 47(C): 94–11 (2014).

[22] Molavi, Pooya, "Simple Models and Biased Forecasts," working paper, Northwestern University, September 2022.

[23] Nagel, Stefan, and Zhengyang Xu, "Asset Pricing with Fading Memory," *Review of Financial Studies* 35: 2190-2245 (2022).

[24] Nakov, Anton, and Galo Nuño, "Learning from Experience in the Stock Market," *Journal of Economic Dynamics and Control* 52: 224-239 (2015).

[25] Neligh, Nathaniel, "Rational Memory with Decay," working paper, University of Tennessee Knoxville, July 2022.

[26] Prat-Carrabin, Arthur, Florent Meyniel, and Rava Azeredo da Silveira, "Resource-Rational Account of Sequential Effects in Human Prediction," *bioRXiv,* posted June 22, 2022. [URL: https://www.biorxiv.org/content/10.1101/2022.06.20.496900v1]

[27] Prat-Carrabin, Arthur, Florent Meyniel, Misha Tsodyks, and Rava Azeredo da Silveira, "Biases and Variability from Costly Bayesian Inference," *Entropy* 23: 603 (2021).

[28] Schraeder, Stefanie, "Information Processing and Non-Bayesian Learning in Financial Markets," *Review of Finance* 20: 823-853 (2016).

[29] Sims, Christopher A., "Implications of Rational Inattention," *Journal of Monetary Economics* 50: 665-690 (2003).

[30] Slobodyan, Sergey, and Raf Wouters, "Learning in an Estimated Medium-scale DSGE Model," *Journal of Economic Dynamics and Control* 36: 26–46 (2012).

[31] Sung, Yeji, "Macroeconomic Expectations and Cognitive Noise," working paper, Columbia University, November 2022.

[32] Yu, Angela J., and Jonathan D. Cohen, "Sequential Effects: Superstition or Rational Behavior?" *Advances in Neural Information Processing Systems* 21: 1873-1880 (2009).

For Online Publication
APPENDIX

Azeredo da Silveira, Sung, and Woodford,
"Optimally Imprecise Memory and Biased Forecasts"

A Reduction of the General Forecasting Problem to Estimation of μ

Consider the problem of choosing the vector of forecasts z_t each period so as to minimize (2.2). The elements of z_t must be chosen as a function of the DM's cognitive state at time t (after observing the external state y_t). As explained in the text, the DM's cognitive state at time t is assumed to consist of the value of the current external state y_t (observed with perfect precision), along with whatever additional information is reflected in the DM's period t memory state m_t. (In this section, it is not yet necessary to specify the nature of the vector m_t.)

If we use the notation $\mathrm{E}_t[\cdot]$ for the expectation of a random variable conditional on a complete description of the state at date t (including knowledge of the true value of μ), then

$$\mathrm{E}[(z_t - \mathrm{E}_t\tilde{z}_t)'W(\tilde{z}_t - \mathrm{E}_t\tilde{z}_t)] \;=\; 0,$$

since $\tilde{z}_t - \mathrm{E}_t\tilde{z}_t$ is a function of innovations in the external state subsequent to date t, that must be distributed independently of all of the determinants of both z_t and $\mathrm{E}_t\tilde{z}_t$. It follows that the term in (2.2) involving z_t can be equivalently expressed as[66]

$$\begin{aligned}\mathrm{E}[(z_t - \tilde{z}_t)'W(z_t - \tilde{z}_t)] \;&=\; \mathrm{E}[(z_t - \mathrm{E}_t\tilde{z}_t)'W(z_t - \mathrm{E}_t\tilde{z}_t)] \\ &\quad +\mathrm{E}[(\tilde{z}_t - \mathrm{E}_t\tilde{z}_t)'W(\tilde{z}_t - \mathrm{E}_t\tilde{z}_t)] \\ &\equiv\; L_{1t} \,+\, L_{2t}.\end{aligned}$$

Moreover, L_{2t} is independent of the decisions of the DM, and thus irrelevant to a determination of the optimal decision rule. The loss function (2.2) can thus equivalently be written as the discounted sum of the L_{1t} terms, which involve squared differences between z_t and $\mathrm{E}_t\tilde{z}_t$.

It further follows from the law of motion (2.1) that

$$\mathrm{E}_t\tilde{z}_t \;=\; \sum_{j=0}^{\infty} A_j[\mu + \rho^j(y_t - \mu)].$$

Since the precise value of y_t is presumed to be part of the cognitive state on the basis of which z_t can be chosen, one can write any decision rule in the form

$$z_t \;=\; \hat{z}_t \,+\, (\sum_{j=0}^{\infty} \rho^j A_j) \cdot y_t,$$

[66] Here we omit the factor β^t that multiplies this term in (2.2).

where $\hat{z}_t$ must be some function of the cognitive state at date t. In terms of this notation, the relevant part of the loss function (2.2) can then be written as

$$L_{1t} \;=\; \mathrm{E}[(\hat{z}_t - \mu a)' W (\hat{z}_t - \mu a)],$$

where we define $a \equiv \sum_{j=0}^{\infty}(1-\rho^j)A_j$ and make use of the fact that $\mathrm{E}_t[\mu] = \mu$.

The term L_{1t} that we wish to minimize can further be expressed as the expected value (integrating over all possible realizations of the cognitive state s_t in period t) of the quantity

$$\begin{aligned}\tilde{L}_1(s_t) \;&\equiv\; \mathrm{E}[(\hat{z}_t - \mu a)' W (\hat{z}_t - \mu a)\,|s_t] \\ &=\; \mathrm{E}[\hat{z}_t\,|s_t]' W \mathrm{E}[\hat{z}_t\,|s_t] \;+\; \mathrm{E}[\breve{z}_t' W \breve{z}_t\,|s_t] \\ &\quad -2a'W\mathrm{E}[\hat{z}_t\,|s_t]\cdot\mathrm{E}[\mu|s_t] \;+\; a'Wa\cdot\mathrm{E}[\mu^2|s_t],\end{aligned}$$

where we define $\breve{z}_t \equiv \hat{z}_t - \mathrm{E}[\hat{z}_t\,|s_t]$. (In expanding the right-hand side in this way, we use the fact that $\mathrm{E}[\breve{z}_t\,|s_t] = 0$, and that $\breve{z}_t$ must be independent of the deviation of μ from $\mathrm{E}[\mu|s_t]$, since the DM has no way to condition her action on μ except through the information about μ revealed by the cognitive state.) The expression $\tilde{L}_1(s_t)$ can then be separately minimized for each possible cognitive state s_t, by choosing a distribution for $\hat{z}_t$ conditional on that state. We further note that the random component $\breve{z}_t$ of the action affects only the second term on the right-hand side, and so should be chosen to minimize that term; since W is positive definite, this is achieved by setting $\breve{z}_t = 0$ with certainty, so that $\hat{z}_t$ must be a deterministic function of s_t.

We can then simply write $\mathrm{E}[\hat{z}_t\,|s_t]$ as $\hat{z}_t$, and observe that

$$\tilde{L}_1(s_t) \;=\; (\hat{z}_t - a\mathrm{E}[\mu|s_t])' W (\hat{z}_t - a\mathrm{E}[\mu|s_t]) \;+\; a'Wa\cdot\mathrm{var}[\mu|s_t], \tag{A.1}$$

where the final term on the right-hand side is independent of the choice of $\hat{z}_t$. Thus in each cognitive state s_t, $\hat{z}_t$ must be chosen to minimize the first term on the right-hand side; since W is positive definite, this is achieved by setting $\hat{z}_t = a\cdot\hat{\mu}_t$, where $\hat{\mu}_t = \mathrm{E}[\mu|s_t]$.

Thus there is no loss of generality in restricting the DM to response rules of the form $\hat{z}_t = a\cdot\hat{\mu}_t$, where $\hat{\mu}_t$ is a scalar choice that depends on the cognitive state in period t, and that can be interpreted as the DM's estimate of μ given the cognitive state. Substituting this expression for $\hat{z}_t$ into (A.1), we have

$$\begin{aligned}\tilde{L}_1(s_t) \;&=\; a'Wa\cdot\left\{(\hat{\mu}_t - \mathrm{E}[\mu|s_t])^2 \;+\; \mathrm{var}[\mu(s_t)]\right\} \\ &=\; a'Wa\cdot\mathrm{E}[(\hat{\mu}_t - \mu)^2\,|s_t].\end{aligned}$$

Then taking the unconditional expectation of this expression, we obtain

$$L_{1t} \;=\; \alpha\cdot MSE_t,$$

where $\alpha \;\equiv\; a'Wa \;>\; 0$ and MSE_t is defined as in the text.

Under any forecasting rule of the kind assumed here, then, the value of the loss function (2.2) will equal (2.4), plus an additional term

$$\sum_{t=0}^{\infty}\beta^t L_{2t}$$

that is independent of the DM's forecasting rule. Hence within this class of forecasting rules, the rule that minimizes (2.2) must be the one that minimizes (2.4); and since any other kind of forecasting rule can only lead to a higher value of (2.2), we can replace the problem of choosing a rule for determining z_t that minimizes (2.2) by the problem of choosing a rule for determining $\hat{\mu}_t$ that minimizes (2.4).

B Bayesian Updating After the External State is Observed: A Kalman Filter

Let the elements of the memory state be partitioned as

$$m_t \;=\; \begin{bmatrix} \underline{m}_t \\ \bar{m}_t \end{bmatrix}, \tag{B.1}$$

where the lower block consists of the elements of the reduced memory state

$$\bar{m}_t \;\equiv\; \mathrm{E}[x_t \,|m_t], \qquad \text{where } \; x_t \;\equiv\; \begin{bmatrix} \mu \\ y_{t-1} \end{bmatrix},$$

while the upper block consists of the conditional expectations $\mathrm{E}[y_{t-j}\,|m_t]$ for $2 \leq j \leq t$. (This simply requires an appropriate ordering of the elements of m_t, using the notation for this vector introduced in the main text.)

We assume a posterior distribution of the form

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

conditional on the memory state m_t, where $\bar{m}_t$ is a 2-vector and Σ_t is a 2×2 symmetric, p.s.d. matrix. Under our assumption of linear-Gaussian dynamics for the memory state, the vector $\bar{m}_t$ will also be drawn from a multivariate Gaussian distribution. Since the prior for the hidden state vector is specified to be

$$x_t \;\sim\; N(0,\, \Sigma_0), \qquad\qquad \Sigma_0 \;\equiv\; \begin{bmatrix} \Omega & \Omega \\ \Omega & \Omega + \sigma_y^2 \end{bmatrix}, \tag{B.2}$$

it follows that the unconditional distribution for the reduced memory state $\bar{m}_t$ must be of the form

$$\bar{m}_t \;\sim\; N(0,\, \Sigma_0 - \Sigma_t).$$

The complete set of variables (x_t, m_t) also have a multivariate Gaussian distribution. Moreover, since (by assumption) the expectation of x_t conditional on the realization of m_t depends only on the elements of $\bar{m}_t$, it follows that the entire distribution of x_t conditional on m_t depends only on $\bar{m}_t$, so that

$$x_t|m_t \;=\; x_t|\bar{m}_t.$$

Hence the joint distribution of the variables (x_t, m_t) can be factored as

$$p(x_t, \underline{m}_t, \bar{m}_t) \;=\; p(x_t, \bar{m}_t) \cdot p(\underline{m}_t\,|\bar{m}_t).$$

The DM then observes the external state y_t, which is assumed to depend on the hidden state vector x_t through an "observation equation" of the form

$$y_t \;=\; c'x_t \,+\, \epsilon_{yt}, \qquad \epsilon_{yt} \,\sim\, N(0,\, \sigma_\epsilon^2)$$

as a consequence of (2.1), where we further assume that ϵ_{yt} is distributed independently of both m_t and x_t. It follows that the variables (x_t, m_t, y_t) will have a joint distribution that is multivariate Gaussian; and that this distribution can be factored as

$$\begin{aligned} p(x_t, m_t, y_t) \;&=\; p(x_t, m_t) \cdot p(y_t\,|x_t) \\ &=\; p(\underline{m}_t\,|\bar{m}_t) \cdot p(x_t, \bar{m}_t) \cdot p(y_t\,|x_t) \\ &=\; p(\underline{m}_t\,|\bar{m}_t) \cdot p(x_t, \bar{m}_t, y_t). \end{aligned}$$

From this it follows that

$$x_t\,|m_t, y_t \;=\; x_t\,|\bar{m}_t, y_t.$$

Thus both the expectation of x_t conditional on the cognitive state $s_t \equiv (m_t, y_t)$, and the variance-covariance matrix of the errors in the estimation of x_t based on the cognitive state, will depend only on the joint distribution of the variables $(x_t, \bar{m}_t, y_t)$. Moreover, the distribution for x_t conditional on the realizations of the elements of the cognitive state will be multivariate Gaussian,

$$x_t\,|\bar{m}_t, y_t \;\sim\; N(\bar{\mu}_t,\, \bar{\Sigma}_t), \tag{B.3}$$

where $\bar{\mu}_t$ is a linear function of $\bar{m}_t$ and y_t, while $\bar{\Sigma}_t$ is independent of the realizations of either $\bar{m}_t$ or y_t.

We can further decompose the vector of means $\bar{\mu}_t$ as

$$\begin{aligned} \bar{\mu}_t \;&=\; \mathrm{E}[x_t\,|\bar{m}_t, y_t] \\ &=\; \mathrm{E}[x_t\,|\bar{m}_t] \,+\, \{\mathrm{E}[x_t|\bar{m}_t, y_t] - \mathrm{E}[x_t|\bar{m}_t]\} \\ &=\; \bar{m}_t \,+\, \gamma_t \cdot (y_t \,-\, \mathrm{E}[y_t\,|\bar{m}_t]) \\ &=\; \bar{m}_t \,+\, \gamma_t \cdot (y_t \,-\, c'\mathrm{E}[x_t\,|\bar{m}_t]) \\ &=\; \bar{m}_t \,+\, \gamma_t \cdot (y_t \,-\, c'\bar{m}_t), \end{aligned}$$

where γ_t is the vector of *Kalman gains.* (The first element of this vector equation is then just equation (2.6) in the main text.)

The vector of Kalman gains must be chosen so that the estimation errors $x_t - \bar{\mu}_t$ are orthogonal to the surprise in the observation of the external state, $y_t - c'\bar{m}_t$. This requires that

$$\begin{aligned} 0 \;&=\; \mathrm{cov}(x_t - \bar{\mu}_t,\, y_t - c'\bar{m}_t) \\ &=\; \mathrm{cov}((x_t - \bar{m}_t) - \gamma_t(y_t - c'\bar{m}_t),\, y_t - c'\bar{m}_t) \\ &=\; \mathrm{var}[x_t - \bar{m}_t]c \,-\, \mathrm{var}[c'(x_t - \bar{m}_t) + \epsilon_{yt}] \cdot \gamma_t \\ &=\; \Sigma_t c \,-\, [c'\Sigma_t c + \sigma_\epsilon^2] \cdot \gamma_t. \end{aligned}$$

Hence

$$\gamma_t \;=\; \frac{\Sigma_t c}{c'\Sigma_t c + \sigma_\epsilon^2}. \tag{B.4}$$

The gain coefficient γ_{1t} in equation (2.6) is just the first element of this vector, $\gamma_{1t} \equiv e_1'\gamma_t$. This together with (B.4) yields the formula (2.8) given in the main text.

The variance-covariance matrix in the conditional distribution (B.3) will be given by

$$\begin{aligned}
\bar{\Sigma}_t &= \mathrm{var}[x_t - \bar{\mu}_t] \;=\; \mathrm{var}[(x_t - \bar{m}_t) - \gamma_t(y_t - c'\bar{m}_t)] \\
&= \mathrm{var}[(I - \gamma_t c')(x_t - \bar{m}_t) \;-\; \gamma_t \epsilon_{yt}] \\
&= (I - \gamma_t c')\Sigma_t(I - \gamma_t c')' \;+\; \sigma_\epsilon^2 \gamma_t \gamma_t' \\
&= \Sigma_t \;-\; 2[c'\Sigma_t c + \sigma_\epsilon^2]\gamma_t\gamma_t' \;+\; [c'\Sigma_t c]\gamma_t\gamma_t' \;+\; \sigma_\epsilon^2\gamma_t\gamma_t' \\
&= \Sigma_t \;-\; [c'\Sigma_t c + \sigma_\epsilon^2]\gamma_t\gamma_t'.
\end{aligned}$$

The remaining uncertainty about the value of μ given the cognitive state, $\hat{\sigma}_t^2$, is then equal to $\bar{\Sigma}_{11,t}$, so that

$$\hat{\sigma}_t^2 \;=\; e_1'\bar{\Sigma}_t e_1 \;=\; e_1'\Sigma_t e_1 \;-\; (c'\Sigma_t c + \sigma_\epsilon^2)(\gamma_{1t})^2,$$

which is just expression (2.7) in the main text.

Substituting expression (B.2) for Σ_0 into this solution, we obtain

$$\begin{aligned}
\hat{\sigma}_0^2 &= \Omega \;-\; \left(\Omega + \sigma_y^2\right) \cdot \left[\frac{\Omega}{\Omega + \sigma_y^2}\right]^2 \\
&= \frac{\Omega\sigma_y^2}{\Omega + \sigma_y^2},
\end{aligned}$$

which is the formula given in (2.9). It remains to be shown that this is an upper bound for $\hat{\sigma}_t^2$. To show this, we observe that

$$\begin{aligned}
\hat{\sigma}_t^2 &= \min_{\beta,\gamma_1} \mathrm{var}[\mu - \beta'\bar{m}_t - \gamma_1 y_t] \\
&\leq \min_{\gamma_1} \mathrm{var}[\mu - \gamma_1 y_t] \\
&\leq \mathrm{var}[\mu \;-\; (\Omega/(\Omega + \sigma_y^2)) \cdot y_t] \\
&= \mathrm{var}[(\sigma_y^2/(\Omega + \sigma_y^2))\mu \;-\; (\Omega/(\Omega + \sigma_y^2))(y_t - \mu)] \\
&= \left(\frac{\sigma_y^2}{\Omega + \sigma_y^2}\right)^2 \mathrm{var}[\mu] \;+\; \left(\frac{\Omega}{\Omega + \sigma_y^2}\right)^2 \mathrm{var}[y_t|\mu] \\
&= \left(\frac{\sigma_y^2}{\Omega + \sigma_y^2}\right)^2 \Omega \;+\; \left(\frac{\Omega}{\Omega + \sigma_y^2}\right)^2 \sigma_y^2 \\
&= \frac{\Omega\sigma_y^2}{\Omega + \sigma_y^2} \;=\; \sigma_0^2.
\end{aligned}$$

This establishes the upper bound (2.9) stated in the main text.

C Demonstration that an Optimal Memory Structure Records Information Only about the Reduced Cognitive State

Let (1.2) be written in the partitioned form

$$\begin{bmatrix} \underline{m}_{t+1} \\ \bar{m}_{t+1} \end{bmatrix} = \begin{bmatrix} \Lambda_{a,t} & \Lambda_{b,t} \\ \Lambda_{c,t} & \Lambda_{d,t} \end{bmatrix} \begin{bmatrix} \underline{s}_t \\ \bar{s}_t \end{bmatrix} + \begin{bmatrix} \underline{\omega}_{t+1} \\ \bar{\omega}_{t+1} \end{bmatrix}. \tag{C.1}$$

Here m_{t+1} is again partitioned as in (B.1). The lower block of s_t consists of the elements of the reduced cognitive state

$$\bar{s}_t \;\equiv\; \begin{bmatrix} \hat{\mu}_t \\ y_t \end{bmatrix},$$

both elements of which are linear functions of s_t, as a consequence of equation (2.6). We choose a representation for the vector s_t such that the lower block consists of the elements of $\bar{s}_t$, the elements of $\underline{s}_t$ are all uncorrelated with the elements of $\bar{s}_t$, and the elements of the vectors $\bar{s}_t$ and $\underline{s}_t$ together span the same linear space of random variables as the elements of s_t. (We can necessarily write any memory structure of the form (1.2) in this way; it amounts simply to a choice of the basis vectors in terms of which the vectors m_{t+1} and s_t are each decomposed.)

Let us suppose furthermore that a representation for m_{t+1} is chosen consistent with the normalization $\mathrm{E}[\bar{s}_t\,|m_{t+1}] \;=\; \bar{m}_{t+1}$. This holds if and only if both elements of the vector $\bar{s}_t - \bar{m}_{t+1}$ are uncorrelated with each of the elements of m_{t+1}. These consistency conditions can be reduced to two requirements: (i) the requirement that

$$\mathrm{var}[\Lambda_{c,t}\underline{s}_t + \bar{\omega}_{t+1}] \;=\; (I - \Lambda_{d,t})X_t\Lambda'_{d,t}, \tag{C.2}$$

where the matrix $X_t \;\equiv\; \mathrm{var}[\bar{s}_t]$ is independent of the memory structure chosen for period t; and (ii) the requirement that $\bar{s}_t - \bar{m}_{t+1}$ be uncorrelated with all elements of $\underline{m}_{t+1}$. (Note that $\bar{s}_t - \bar{m}_{t+1}$ is uncorrelated with $\bar{m}_{t+1}$ if and only if (C.2) holds.)

C.1 Forecast accuracy depends only on the matrices $\{\Lambda_{d,t}\}$

Suppose that in any period t, we take the memory structure in periods $\tau < t$ as given. This means that the DM's uncertainty about x_t given the memory state m_t (specified by the posterior variance-covariance matrix Σ_t) will be given. (If $t = 0$, Σ_0 is simply given by the prior.) Hence the value of $\hat{\mu}_t$ as a function of $\bar{m}_t$ and y_t will be given, and consequently the value of MSE_t will be given, following the discussion in the main text (and the previous section of this appendix). The elements of the matrix X_t will similarly be given.

We next consider how $\Lambda_{d,t}$ must be chosen, in order for it to be possible to choose matrices $\Lambda_{c,t}$ and $\mathrm{var}[\bar{\omega}_{t+1}]$ such that (C.2) is satisfied. Equation (C.2) requires that $(I - \Lambda_{d,t})X_t\Lambda'_{d,t}$, be a symmetric matrix; this will hold if and only if the simpler requirement is satisfied that $\Lambda_{d,t}X_t = X_t\Lambda'_{d,t}$ be a symmetric matrix. In addition, it is necessary that $(I - \Lambda_{d,t})X_t\Lambda'_{d,t}$ be a p.s.d. matrix. The set of matrices $\Lambda_{d,t}$ with these properties is a non-empty set ($\Lambda_{d,t} = 0$ is a trivial example), and depends only on the matrix X_t. Let this set of matrices be denoted $\mathcal{L}(X_t)$.

Now let $\Lambda_{d,t}$ be any matrix that belongs to $\mathcal{L}(X_t)$. Then it is possible to choose the matrices $\Lambda_{c,t}$ and $\mathrm{var}[\bar{\omega}_{t+1}]$ so that (C.2) is satisfied; and given any such choice of these two matrices, it is further possible to choose the specification of the equation for $\underline{m}_{t+1}$ so that all elements of $\underline{m}_{t+1}$ are uncorrelated with the elements of $\bar{s}_t - \bar{m}_{t+1}$. Given any such specifications, both conditions (i) and (ii) above will be satisfied. Thus the matrix $\Lambda_{d,t}$ is admissible as part of the specification of a memory structure; and any possible memory structure consistent with the matrix $\Lambda_{d,t}$ will be one of those with the properties just assumed.

Given a matrix $\Lambda_{d,t}$ of this sort, we next observe that the equations determining $\bar{m}_{t+1}$ can be written in the form

$$\bar{m}_{t+1} \;=\; \Lambda_{d,t}\bar{s}_t \;+\; \nu_{t+1},$$

where $\nu_{t+1} \sim N(0,\, \Lambda_{d,t}X_t)$ is distributed independently of $\bar{s}_t$. Thus the joint distribution of $(\bar{s}_t, \bar{m}_{t+1})$ will be a multivariate Gaussian distribution, the parameters of which are completely determined by X_t and $\Lambda_{d,t}$. It then follows that the conditional distribution $\bar{s}_t|\bar{m}_{t+1}$ will be a bivariate Gaussian distribution, with a mean $\bar{m}_{t+1}$ and a variance independent of the realization of $\bar{m}_{t+1}$, which also depends only on X_t and $\Lambda_{d,t}$. Moreover, since the elements of $\underline{m}_{t+1}$ are all Gaussian random variables distributed independently of $\bar{s}_t - \bar{m}_{t+1}$, knowledge of $\underline{m}_{t+1}$ cannot further improve one's estimate of $\bar{s}_t$, and so the conditional distribution $\bar{s}_t|m_{t+1} = \bar{s}_t|\bar{m}_{t+1}$. Finally, since we can write

$$x_{t+1} \;=\; \bar{s}_t \;+\; \begin{bmatrix} u_t \\ 0 \end{bmatrix},$$

where $u_t \sim N(0,\, \hat{\sigma}_t^2)$ must be uncorrelated with any of the elements of s_t (and hence uncorrelated with any of the elements of m_{t+1}), we must further have

$$x_{t+1}|m_{t+1} \;\sim\; N(\bar{m}_{t+1},\, \Sigma_{t+1})$$

where

$$\Sigma_{t+1} \;=\; \mathrm{var}[\bar{s}_t\,|\bar{m}_{t+1}] \;+\; \hat{\sigma}_t^2\, e_1 e_1'.$$

Since $\hat{\sigma}_t^2$ also depends only on Σ_t (see equation (2.7)), it follows that the elements of Σ_{t+1} depend only on Σ_t and $\Lambda_{d,t}$.

This argument can then be used recursively (starting from period $t = 0$) to show that given the initial uncertainty matrix Σ_0 implied by the prior (B.2), we can completely determine the entire sequence of matrices $\{\Sigma_t\}$, given a sequence of matrices $\{\Lambda_{d,t}\}$ for all $t \geq 0$ with the property that for each t, $\Lambda_{d,t} \in \mathcal{L}(X_t)$, where X_t is the matrix implied by Σ_t. Moreover, given such a sequence of matrices $\{\Lambda_{d,t}\}$, the value of MSE_t for each period t will be uniquely determined as well. Hence the terms in the loss function (2.5) that depend on the accuracy of forecasts that are possible using a given memory structure will depend only on the sequence of matrices $\{\Lambda_{d,t}\}$. (These matrices must be chosen to satisfy a set of consistency conditions, stated above, but these conditions can also be expressed purely in terms of the sequence of matrices $\{\Lambda_{d,t}\}$.) Thus the other elements of the specification (C.1) of the memory structure matter only to the extent that they have consequences for the information cost terms in (2.5).

C.2 Mutual information: a useful lemma

Information costs in period t are assumed to be an increasing function of $I_t = I(M;S)$, the Shannon mutual information between random variables M (the realizations of which are denoted m_{t+1}) and S (the realizations of which are denoted s_t).[67] Each of the random vectors M and S can further be partitioned as $M = (\underline{M}, \bar{M})$, $S = (\underline{S}, \bar{S})$.

Now for any random variables $X_1, X_2, \ldots,$ let $H(X_1, X_2, \ldots, X_k)$ be the entropy of the joint distribution for variables $(X_1, X_2, \ldots, X_k)$, and $H(X_1, \ldots, X_k \,|X_{k+1}, \ldots X_{k+m})$ be the entropy of the joint distribution of the variables $(X_1, \ldots, X_k)$ conditional on the values of the variables $(X_{k+1}, \ldots X_{k+m})$. The chain rule for entropy implies that

$$H(X_1, X_2, \ldots, X_k) \;=\; H(X_1) \,+\, H(X_2\,|X_1) \,+\, \ldots \,+\, H(X_k\,|X_1, \ldots, X_{k-1}).$$

We can then define the mutual information between the variables $(X_1, \ldots, X_k)$ and the variables $(X_{k+1}, \ldots X_{k+m})$ as

$$I(X_1, \ldots, X_k;\, X_{k+1}, \ldots, X_{k+m}) \;\equiv\; H(X_1, \ldots, X_k) \,-\, H(X_1, \ldots, X_k\,|X_{k+1}, \ldots X_{k+m}).$$

(The information about the first set of variables that is revealed by learning the values of the second set of variables is measured by the average amount by which the entropy of the conditional distribution is smaller than the entropy of the unconditional distribution of the first set of variables.) Similarly, we can define the mutual information between the first set of variables and the second set of variables, conditioning on the values of some third set of variables as

$$\begin{aligned} &I(X_1, \ldots, X_k;\, X_{k+1}, \ldots, X_{k+m}\,|X_{k+m+1}, \ldots, X_{k+m+n}) \\ &\qquad \equiv\; H(X_1, X_2, \ldots, X_k\,|X_{k+m+1}, \ldots, X_{k+m+n}) \,-\, H(X_1, \ldots, X_k\,|X_{k+1}, \ldots, X_{k+m+n}). \end{aligned}$$

Thus for any set of four random variables $\underline{M}, \bar{M}, \underline{S}, \bar{S}$, we must have

$$\begin{aligned} I(\underline{S}, \bar{S};\, \underline{M}, \bar{M}) \;&=\; H(\underline{S}, \bar{S}) \,-\, H(\underline{S}, \bar{S}\,|\underline{M}, \bar{M}) \\ &=\; [H(\bar{S}) + H(\underline{S}\,|\bar{S})] \,-\, [H(\bar{S}\,|\underline{M}, \bar{M}) + H(\underline{S}\,|\bar{S}, \underline{M}, \bar{M})] \\ &=\; [H(\bar{S}) + H(\underline{S}\,|\bar{S})] \,-\, [H(\bar{S}, \underline{M}, \bar{M}) - H(\underline{M}\,|\bar{M}) - H(\bar{M})] \,-\, H(\underline{S}\,|\bar{S}, \underline{M}, \bar{M}) \\ &=\; [H(\bar{S}) + H(\underline{S}\,|\bar{S})] \,-\, [(H(\bar{M}) + H(\bar{S}\,|\bar{M}) + H(\underline{M}\,|\bar{M}, \bar{S})) - H(\underline{M}\,|\bar{M}) - H(\bar{M})] \\ &\qquad\qquad\qquad\qquad\qquad - H(\underline{S}\,|\bar{S}, \underline{M}, \bar{M}) \\ &=\; [H(\bar{S}) + H(\underline{S}\,|\bar{S})] \,-\, [H(\bar{S}\,|\bar{M}) + H(\underline{M}\,|\bar{M}, \bar{S}) - H(\underline{M}\,|\bar{M})] \,-\, H(\underline{S}\,|\bar{S}, \underline{M}, \bar{M}) \\ &=\; [H(\bar{S}) - H(\bar{S}\,|\bar{M})] \,+\, [H(\underline{S}\,|\bar{S}) - H(\underline{S}\,|\bar{S}, \underline{M}, \bar{M})] \,+\, [H(\underline{M}\,|\bar{M}) - H(\underline{M}\,|\bar{M}, \bar{S})] \\ &=\; I(\bar{S};\, \bar{M}) \,+\, I(\underline{S};\, \underline{M}, \bar{M}\,|\bar{S}) \,+\, I(\underline{M};\, \bar{S}\,|\bar{M}). \end{aligned}$$

Then, since mutual information is necessarily non-negative, we can establish the lower bound

$$I_t \;=\; I(\underline{S}, \bar{S};\, \underline{M}, \bar{M}) \;\geq\; I(\bar{S};\, \bar{M}). \tag{C.3}$$

[67] Here we adopt the notation used in Cover (2006), with different symbols for the random variables M and S and their realizations. This is to make it clear that I_t is not a function of the values taken by m_{t+1} and s_t along a particular history, but instead a function of the complete joint distribution of the two random variables; I_t is itself not a random variable, but a single number for each date t.

Furthermore, this lower bound is achieved if and only if

$$I(\underline{S};\, \underline{M}, \bar{M}\,|\bar{S}) \;=\; I(\underline{M};\, \bar{S}\,|\bar{M}) \;=\; 0.$$

For any three random variables X, Y, Z, the conditional mutual information $I(X;\, Y\,|Z) = 0$ if and only if the variables X and Y are distributed independently one another, conditional on the value of Z. Hence the lower bound (C.3) is achieved if and only if (a) conditional on the value of $\bar{m}_{t+1}$, the variables $\bar{s}_t$ and $\underline{m}_{t+1}$ are independent of one another; and (b) conditional on the value of $\bar{s}_t$, the variables $\underline{s}_t$ and m_{t+1} are independent of one another.

C.3 Optimality of Setting $\Lambda_{a,t} = \Lambda_{b,t} = \Lambda_{c,t} = 0$

We return now to the consideration of possible memory structures. Let the sequence of matrices $\{\Lambda_{d,t}\}$ be chosen to satisfy the consistency conditions discussed above, and for a given such sequence, consider an optimal choice of the remaining elements of the specification (C.1), from among those specifications that are consistent with the sequence $\{\Lambda_{d,t}\}$ (that is, that will satisfy both conditions (i) and (ii) stated above).

We have shown above that the sequence of values $\{MSE_t\}$ is completely determined by the specification of $\{\Lambda_{d,t}\}$. Hence other aspects of the specification of the memory structure can matter only to the extent that they affect the sequence of values $\{I_t\}$. Moreover, we have shown that the joint distribution of $(\bar{s}_t, \bar{m}_{t+1})$ each period is completely determined by X_t and $\Lambda_{d,t}$, which means that the lower bound for I_t given in (C.3) is completely determined by the choice of $\{\Lambda_{d,\tau}\}$ for $\tau \leq t$. It thus remains only to consider whether this lower bound can be achieved, and under what conditions.

We first observe that the lower bound is achievable. For any sequence of matrices $\{\Lambda_{d,t}\}$ satisfying the specified conditions, a memory structure specification with $\Lambda_{a,t} = \Lambda_{b,t} = \Lambda_{c,t} = 0$, together with a stipulation that $\underline{\omega}_{t+1}$ be distributed independently of $\bar{\omega}_{t+1}$ and that $\mathrm{var}[\bar{\omega}_{t+1}] = \Lambda_{d,t}X_t$, will satisfy both conditions (i) and (ii) stated in the introduction to this appendix, and thus this represents a feasible memory structure. One can also show that such a specification satisfies both of conditions (a) and (b) stated at the end of section C.2, so that the lower bound (C.3) is achieved in each period. Thus such a specification achieves the lowest possible value for the combined objective function (2.5), and will be optimal, given our choice of the sequence $\{\Lambda_{d,t}\}$.

Not only will this specification be sufficient for achieving the lowest possible value of (2.5), but it will be essentially necessary. We have shown above that achieving the lower bound for I_t in period t requires that conditional on the value of $\bar{s}_t$, the variables $\underline{s}_t$ and m_{t+1} are independent of one another. This means that the values of the variables in the vector $\underline{s}_t$ cannot help at all in predicting any elements of m_{t+1}, once one is already using the reduced cognitive state $\bar{s}_t$ to forecast the next period's memory state; thus one must be able to write law of motion (C.1) for the memory state with $\Lambda_{a,t} = \Lambda_{c,t} = 0$.[68] Thus it is necessarily the

[68] It might be possible to satisfy the condition required for the lower bound with non-zero elements in one of these matrices; but this will occur only because of collinearity in the fluctuations in the elements of the vector $\underline{s}_t$, so that it is possible to have a law of motion in which $\underline{s}_t$ has no effect on m_{t+1}, despite non-zero matrices $\Lambda_{a,t}$ and $\Lambda_{c,t}$. In such a case, the representation of the cognitive state by the vector s_t would involve redundancy; and in any event, there would be no loss of generality in setting $\Lambda_{a,t} = \Lambda_{c,t} = 0$, since the implied fluctuations in the memory state would be the same.

case that the elements of m_{t+1} convey information only about the reduced cognitive state $\bar{s}_t$, and not about any other aspects of the cognitive state s_t.

In addition, we have shown above that achieving the lower bound for I_t in period t requires that conditional on the value of $\bar{m}_{t+1}$, the variables $\bar{s}_t$ and $\underline{m}_{t+1}$ are independent of one another. Thus all of the information about $\bar{s}_t$ that is contained in the memory state m_{t+1} is contained in the elements $\bar{m}_{t+1}$. This means either that $\Lambda_{b,t} = 0$ as well, or, to the extent that some element of $\underline{m}_{t+1}$ corresponds to a row of $\Lambda_{b,t}$ with non-zero elements, that element of $\underline{m}_{t+1}$ must be a linear combination of the elements of $\bar{m}_{t+1}$, so that conditioning upon its value conveys no new information about $\bar{s}_t$. Thus any specification of the memory structure in which $\Lambda_{b,t} \neq 0$ in any period represents a redundant representation of the contents of memory available in period $t+1$; we can equivalently describe the contents of memory by eliminating all such rows from m_{t+1}.

Thus there is no loss of generality in assuming that the lower bound is achieved by specifying $\Lambda_{a,t} = \Lambda_{b,t} = \Lambda_{c,t} = 0$ in each period. Finally, satisfaction of consistency condition (ii) in this case requires that the elements of $\underline{\omega}_{t+1}$ be distributed independently of the elements of $\bar{\omega}_{t+1}$. We might still allow var$[\underline{\omega}_{t+1}]$ to be non-zero; this would mean that $\underline{m}_{t+1}$ contains elements that fluctuate randomly, but are completely uncorrelated with the previous period's cognitive state s_t. Such an information structure is equally optimal, in the sense that (2.5) is made no larger by the existence of such components of the memory state, given our assumption that only mutual information is costly. But the additional components $\underline{m}_{t+1}$ of the memory structure will have no consequences for cognitive processing, and our inclusion of them as part of the representation of the memory state violates our assumption in the text that we label memory states by their implied posteriors for the values of μ and the past realizations of the external state; using labels $(\underline{m}_{t+1}, \bar{m}_{t+1})$ in which $\underline{m}_{t+1}$ is non-null will mean having separate labels for memory states that imply the same posterior (since the value of $\underline{m}_{t+1}$ would be completely uninformative about either μ or any past external states).

Hence in the case of any optimal memory structure, the memory state can be described more compactly by identifying it with the reduced memory state $\bar{m}_{t+1}$, which evolves according to

$$\bar{m}_{t+1} \; = \; \bar{\Lambda}_t \bar{s}_t \, + \, \bar{\omega}_{t+1}, \tag{C.4}$$

where $\bar{\Lambda}_t$ is the matrix called $\Lambda_{d,t}$ in (C.1). (This corresponds to equation (2.12) in the main text.) We need only consider (at most) a two-dimensional memory state, and the optimal memory state conveys information only about the reduced cognitive state $\bar{s}_t$, not about any other aspects of the cognitive state s_t.

C.4 An alternative representation for the reduced cognitive state

We have shown in the main text (equation (2.15)) that the variance matrix of the reduced cognitive state $\bar{s}_t$ can be written as a function of the single parameter $\hat{\sigma}_t^2$:

$$X_t \; = \; X(\hat{\sigma}_t^2) \; \equiv \; \left[\begin{array}{cc} \Omega - \hat{\sigma}_t^2 & \Omega \\ \Omega & \Omega + \sigma_y^2 \end{array} \right].$$

There is another way of writing this function that will be useful below.

We can orthogonalize the reduced cognitive state using the transformation $\bar{s}_t = \Gamma \check{s}_t$, where

$$\Gamma \equiv \begin{bmatrix} 1 & \frac{\Omega}{\Omega+\sigma_y^2} \\ 0 & 1 \end{bmatrix}. \tag{C.5}$$

The elements of the orthogonalized cognitive state have the interpretation

$$\check{s}_t \equiv \begin{bmatrix} \hat{\mu}_t - \mathrm{E}[\mu|y_t] \\ y_t \end{bmatrix},$$

from which it is obvious that the first element must be uncorrelated with the second.

The variance matrix of $\check{s}_t$ is therefore diagonal:

$$\mathrm{var}[\check{s}_t] = \check{X}(\hat{\sigma}_t^2) \equiv \begin{bmatrix} \hat{\sigma}_0^2 - \hat{\sigma}_t^2 & 0 \\ 0 & \Omega + \sigma_y^2 \end{bmatrix}. \tag{C.6}$$

We can then alternatively write

$$X(\hat{\sigma}_t^2) = \Gamma \check{X}(\hat{\sigma}_t^2)\Gamma'. \tag{C.7}$$

D The Law of Motion for the Memory State and the Information Content of Memory

We now consider how the parameterization of the law of motion (C.4) for the memory state determines the degree of uncertainty about the external state vector that will exist when beliefs are conditioned on the memory state, and how the same parameters determine the mutual information between the memory state and the prior cognitive state, and hence the size of the information cost term $c(I_t)$.

We begin by recapitulating the conditions that the sequence of matrices $\{\bar{\Lambda}_t\}$ and $\{\Sigma_{\bar{\omega},t+1}\}$ must satisfy, in order for (C.4) to represent a memory structure consistent with the normalization according to which $\mathrm{E}[x_{t+1}\,|\bar{m}_{t+1}] = \bar{m}_{t+1}$. Condition (C.2) will be satisfied if and only if

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

In order for there to be a symmetric, p.s.d. matrix $\Sigma_{\bar{\omega},t+1}$ that satisfies (D.1), it must be the case that $\bar{\Lambda}_t \in \mathcal{L}(X_t)$. As explained above, this means that $\bar{\Lambda}_t X_t = X_t\bar{\Lambda}_t'$ must be a symmetric matrix, and in addition that $(I - \bar{\Lambda}_t)X_t\bar{\Lambda}_t'$ is p.s.d. Note that since

$$X_t\bar{\Lambda}_t' = (I - \bar{\Lambda}_t)X_t\bar{\Lambda}_t' + \bar{\Lambda}_t X_t\bar{\Lambda}_t',$$

and X_t is necessarily a p.s.d. matrix, it follows from the assumption that $(I - \bar{\Lambda}_t)X_t\bar{\Lambda}_t'$ is p.s.d. that $\bar{\Lambda}_t X_t = X_t\bar{\Lambda}_t'$ will also be a p.s.d. matrix; but this latter condition is weaker than the one assumed in our definition of the set $\mathcal{L}(X_t)$. This constitutes the complete set of conditions that must be satisfied for (C.4) to represent a memory structure consistent with our proposed normalization of the vector m_{t+1}.

We can further specialize these conditions in the case that $\bar{\Lambda}_t$ is a singular matrix. (Here we assume that X_t is of full rank.) If $\bar{\Lambda}_t$ is of rank one (or less), it can be written in the form $\bar{\Lambda}_t = u_t v_t'$, where we are furthermore free to normalize the vector v_t' so that $v_t' X_t v_t = 1$. Then the condition that $\bar{\Lambda}_t X_t = X_t \bar{\Lambda}_t'$ will hold only if $u_t(v_t' X_t) = (X_t v_t) u_t'$. This means that u_t must be collinear with $X_t v_t$, so that we must be able to write $u_t = \lambda_t X_t v_t$, for some scalar λ_t. Thus in the singular case, we must be able to write

$$\bar{\Lambda}_t \;=\; \lambda_t X_t v_t v_t', \tag{D.2}$$

where λ_t is a scalar and v_t is a vector such that $v_t' X_t v_t = 1$. Then

$$(I - \bar{\Lambda}_t) X_t \bar{\Lambda}_t' \;=\; \lambda_t (1 - \lambda_t)(X_t v_t)(X_t v_t)'$$

will be a p.s.d. matrix if and only if in addition $0 \leq \lambda_t \leq 1$. Thus a singular matrix $\bar{\Lambda}_t$ is an element of $\mathcal{L}(X_t)$ if and only if it is of the form (D.2) with $0 \leq \lambda_t \leq 1$ and v_t a vector such that $v_t' X_t v_t = 1$.

Consistency with the proposed normalization of m_{t+1} then further requires that

$$\Sigma_{\bar{\omega},t+1} \;=\; \lambda_t (1 - \lambda_t) X_t v_t v_t' X_t. \tag{D.3}$$

This implies that $\Sigma_{\bar{\omega},t+1}$ is a singular matrix; the random vector $\bar{\omega}_{t+1}$ can be written as $\bar{\omega}_{t+1} = X_t v_t \cdot \tilde{\omega}_{t+1}$, where $\tilde{\omega}_{t+1}$ is a scalar random variable, with distribution $N(0,\, \lambda_t(1-\lambda_t)$. It follows that in such a case, the memory state can be given a one-dimensional representation, writing $\bar{m}_{t+1} = X_t v_t \cdot \tilde{m}_{t+1}$, where the scalar memory state $\tilde{m}_{t+1}$ has a law of motion

$$\tilde{m}_{t+1} \;=\; \lambda_t v_t' \bar{s}_t + \tilde{\omega}_{t+1}, \qquad \tilde{\omega}_{t+1} \sim N(0,\, \lambda_t(1 - \lambda_t)). \tag{D.4}$$

In the case that $X_t = X_0$ (the only case in which it is possible for $X_t = X(\hat{\sigma}_t^2)$ to be singular), we have defined $\mathcal{L}(X_0)$ to include only matrices of the special form (2.16) with $0 \leq \lambda_t \leq 1$. In this case, $\bar{\Lambda}_t$ is necessarily of the form (D.2), with the vector v_t given by (2.25). Hence our comments above about the case in which $\bar{\Lambda}_t$ is singular apply also in the case in which X_t is singular, except that in this latter case we have the further restriction that v_t must be given by (2.25). In this special case, (D.3) reduces to

$$\Sigma_{\bar{\omega},t+1} \;=\; \lambda_t (1 - \lambda_t)[\Omega + \sigma_y^2]\, w w'.$$

D.1 The degree of uncertainty implied by a given memory structure

We turn now to the question of how the posterior uncertainty Σ_{t+1} in the following period is determined by the law of motion for the memory state $\bar{m}_{t+1}$ that can be accessed at that time. Note that the variance of the marginal distribution for x_{t+1} can be decomposed as

$$\mathrm{var}[x_{t+1}] \;=\; \mathrm{E}[\mathrm{var}[x_{t+1}\,|m_{t+1}]] \,+\, \mathrm{var}[\mathrm{E}[x_{t+1}\,|m_{t+1}]],$$

where in the first term on the right-hand side, the variance refers to the distribution of values for x_{t+1} conditional on the realization of m_{t+1}, and the expectation is over realizations of m_{t+1}, while in the second term the variance refers to the distribution of values for m_{t+1}, and the expectation is over values of x_{t+1} conditional on the realization of m_{t+1}. Since the marginal distribution for x_{t+1} is the same for all t, and coincides with the prior distribution for x_0 specified in (B.2), the left-hand side must equal the matrix Σ_0 defined there. Hence the variance decomposition can be written as

$$\Sigma_0 \;=\; \Sigma_{t+1} \,+\, \mathrm{var}[\bar{m}_{t+1}],$$

which implies that in any period,

$$\Sigma_{t+1} \;=\; \Sigma_0 \,-\, \mathrm{var}[\bar{m}_{t+1}].$$

Thus in order to understand how the choice of $\bar{\Lambda}_t$ determines Σ_{t+1}, it suffices that we determine the implications for the degree of variation in $\bar{m}_{t+1}$.

A law of motion of the form (C.4) implies that

$$\begin{aligned} \mathrm{var}[\bar{m}_{t+1}] &= \bar{\Lambda}_t X_t \bar{\Lambda}_t' \,+\, \Sigma_{\bar{\omega},t+1} \\ &= \bar{\Lambda}_t X_t \bar{\Lambda}_t' \,+\, (I - \bar{\Lambda}_t) X_t \bar{\Lambda}_t' \\ &= X_t \bar{\Lambda}_t', \end{aligned}$$

where the second line uses (D.1). Hence we obtain the prediction that

$$\Sigma_{t+1} \;=\; \Sigma_0 \,-\, X_t \bar{\Lambda}_t'. \tag{D.5}$$

Note that for any $\bar{\Lambda}_t \in \mathcal{L}(X_t)$, this must be a symmetric, p.s.d. matrix.

Hence for any value of $\hat{\sigma}_t^2$ satisfying $0 \leq \hat{\sigma}_t^2 \leq \hat{\sigma}_0^2$ and any transition matrix $\bar{\Lambda}_t \in \mathcal{L}(X(\hat{\sigma}_t^2))$, we can substitute $X_t = X(\hat{\sigma}_t^2)$ and the value of Σ_{t+1} given by (D.5) into (2.7) to obtain a solution for $\hat{\sigma}_{t+1}^2$ as a function of $\hat{\sigma}_t^2$ and $\bar{\Lambda}_t$. This defines the function $f(\hat{\sigma}_t^2, \bar{\Lambda}_t)$ referred to in the main text. We can then define $\mathcal{L}^{seq}$ as the set of sequences of transition matrices $\{\bar{\Lambda}_t\}$ for all $t \geq 0$ such that

$$\bar{\Lambda}_0 \in \mathcal{L}(X_0), \qquad \bar{\Lambda}_1 \in \mathcal{L}(X(f(\hat{\sigma}_0^2, \bar{\Lambda}_0))), \qquad \bar{\Lambda}_2 \in \mathcal{L}(X(f(f(\hat{\sigma}_0^2, \bar{\Lambda}_0), \bar{\Lambda}_1))),$$

and so on.

Then given any sequence of transition matrices $\{\bar{\Lambda}_t\} \in \mathcal{L}^{seq}$, there will be uniquely defined sequences $\{\hat{\sigma}_t^2, X_t\}$ for all $t \geq 0$. Equation (D.5), together with (B.2), can then be used to uniquely define the implied sequence of matrices $\{\Sigma_t\}$ for all $t \geq 0$. These matrices can in turn be used in (2.8) to define the Kalman gain γ_{1t} for each $t \geq 0$. Thus for any sequence of transition matrices $\{\bar{\Lambda}_t\} \in \mathcal{L}^{seq}$, there will be uniquely determined sequences $\{\Sigma_t, \gamma_{1t}, \hat{\sigma}_t^2, X_t\}$, as stated in the text. These in turn will imply a uniquely determined sequence of losses $\{MSE_t\}$ from forecast inaccuracy, using (2.10).

D.2 The mutual information implied by a given memory structure

Finally, we compute the mutual information I_t in the case that the memory state consists only of a reduced memory state $\bar{m}_{t+1}$, with law of motion (C.4). We first review the definition of mutual information in the case of continuously distributed random variables.

Let X and Y be two random variables, each parameterized using a finite system of coordinates (so that realizations x and y are each represented by finite-dimensional vectors), and suppose that at least Y has a continuous distribution, with a density function $p(y|x)$ such that $p(y|x) > 0$ for all y in the support of Y and all x in the support of X. Suppose also that the marginal distribution for Y can be characterized by a density function $p(y) = \mathrm{E}[p(Y|x)]$, where the expectation is over possible realizations of x, and $p(y) > 0$ for all y in the support of Y. Then we can measure the degree to which knowing the realization of x changes the distribution that one can expect y to be drawn from by the Kullback-Liebler divergence (or relative entropy) of the conditional distribution $p(y|x)$ relative to the marginal distribution $p(y)$, defined as

$$D_{KL}(p(\cdot|x)||p(\cdot)) \;\equiv\; \mathrm{E}\left[\log \frac{p(y|x)}{p(y)}\right] \;\geq\; 0, \tag{D.6}$$

where the expectation is over possible realizations of y, and this quantity is a function of the particular realization x.[69] The *mutual information* $I(X; Y)$ can then be defined as the mean value of this expression,

$$I(X; Y) \;\equiv\; \mathrm{E}[D_{KL}(p(\cdot|x)||p(\cdot))], \tag{D.7}$$

where the expectation is now over possible realization of x, and the mutual information is also necessarily non-negative.[70]

This definition of the mutual information has the attractive feature of being independent of the coordinates used to parameterize the realizations of the variable Y. Suppose that we write $y = \phi(z)$, where $\phi(\cdot)$ is an invertible smooth coordinate transformation between two Euclidean spaces of the same dimension. Then corresponding to the conditional density $p(y|x)$ for any x, there will be a corresponding density function $\tilde{p}(z|x)$ for the random variable Z (which is just the variable Y described using the alternative coordinate system), such that $\tilde{p}(z|x) = p(\phi(z)|x) \cdot D\phi(z)$ for each z, where $D\phi(z)$ is the Jacobian matrix of the coordinate transformation, evaluated at z. It follows that for any z in the support of Z and any x in the support of X,

$$\frac{p(\phi(z)|x)}{p(\phi(z))} \;=\; \frac{\tilde{p}(z|x)}{\tilde{p}(z)},$$

so that

$$D_{KL}(p(\cdot|x)||p(\cdot)) \;=\; D_{KL}(\tilde{p}(\cdot|x)||\tilde{p}(\cdot))$$

for all x. We thus find that the mutual information $I(X; Y)$ will be the same as $I(X; Z)$: it is unaffected by a change in the coordinates used to parameterize Y.[71]

We can similarly define the mutual information in a case in which the support of Y is not the entire Euclidean space, because of the existence of redundant coordinates in the parameterization of realizations y. Suppose that all vectors y in the support of Y are of the form $y = \phi(z)$, where $\phi(\cdot)$ is a smooth embedding of some lower-dimensional Euclidean space (the support of Z) into a higher-dimensional Euclidean space. Then the information about

[69] The value of this quantity is necessarily non-negative because of Jensen's inequality, owing to the concavity of the logarithm.

[70] Note that this definition — rather than the one often given in terms of the average reduction in the entropy of Y from observing X — has the advantage of remaining well-defined even when the random variable Y has a continuous distribution. See Cover and Thomas (2006) for further discussion.

[71] It is equally unaffected by a change in the coordinates used to parameterize X, though we need not show this here.

the possible realizations of y contained in a realization of x is given by the information that x contains about the possible realizations of z. If the joint distribution of X and Z is such that we can define conditional density functions $\tilde{p}(z|x)$, with $\tilde{p}(z|x) > 0$ for all z and x, and a marginal density function $\tilde{p}(z) > 0$ for all z, then we can define the mutual information between X and Z using (D.7) as above. Since mutual information should be independent of the coordinates used to parameterize the variables, we can use the value of $I(X;\, Z)$ as our definition of $I(X;\, Y)$ in this case as well (even though expression (D.6) is not defined in this case).

In the case of interest in this paper, X and Y are variables with a joint distribution that is multivariate Gaussian. Let us consider first the generic case in which the conditional variance-covariance matrix $\mathrm{var}[Y|x]$ is of full rank. (Note that this matrix will be independent of the realization of x, and so can be written $\mathrm{var}[Y|X]$, to emphasize that only the parameters of the joint distribution matter.) In this case $\mathrm{var}[Y]$ is of full rank as well, and for any x and y, the ratio of the density functions satisfies

$$\begin{aligned}\log \frac{p(y|x)}{p(y)} \;&=\; -\frac{1}{2}\log\frac{\det(\mathrm{var}[Y|x])}{\det(\mathrm{var}[Y])} \\ &\quad -\frac{1}{2}(y-\mathrm{E}[y|x])'\mathrm{var}[Y|x]^{-1}(y-\mathrm{E}[y|x]) \;+\; \frac{1}{2}(y-\mathrm{E}[y])'\mathrm{var}[Y]^{-1}(y-\mathrm{E}[y]).\end{aligned}$$

Hence for any x, we have

$$D_{KL}(x) \;=\; -\frac{1}{2}\log\frac{\det(\mathrm{var}[Y|x])}{\det(\mathrm{var}[Y])},$$

and since this will be independent of the realization of x, we similarly will have

$$I(X;\, Y) \;=\; -\frac{1}{2}\log\frac{\det(\mathrm{var}[Y|X])}{\det(\mathrm{var}[Y])}. \tag{D.8}$$

One case in which $\mathrm{var}[Y|x]$ will not be of full rank is if $y = Uz$ for some matrix U, where z is a random vector of lower dimension than that of y. (In this case, the rank of $\mathrm{var}[Y|x]$ cannot be greater than the rank of $\mathrm{var}[Z|x]$, which is at most the dimension of z.) Let us suppose that the rank of U is equal to the dimension of z, so that any vector $y = Uz$ is associated with exactly one vector z. In such a case we can, as discussed above, define the mutual information between X and Y to equal the mutual information between X and Z. If $\mathrm{var}[Z|x]$ is of full rank, then we can use the calculations of the previous paragraph to show that

$$I(X;\, Y) \;=\; I(X;\, Z) \;=\; -\frac{1}{2}\log\frac{\det(\mathrm{var}[Z|X])}{\det(\mathrm{var}[Z])}. \tag{D.9}$$

We turn now to the calculation of the mutual information between the reduced cognitive state $\bar{s}_t$ and the memory state $\bar{m}_{t+1}$, in the case of a law of motion of the form (C.4) for the memory state. We first consider the case in which X_t is of full rank (which, as noted in the text, will be true except when the memory state m_t is completely uninformative). If $\bar{\Lambda}_t$ and $I - \bar{\Lambda}_t$ are also both matrices of full rank, then

$$\mathrm{var}[\bar{m}_{t+1}\,|\bar{s}_t] \;=\; \Sigma_{\bar{\omega},t+1} \;=\; (I - \bar{\Lambda}_t)X_t\bar{\Lambda}_t'$$

will be of full rank, and

$$\mathrm{var}[\bar{m}_{t+1}] \;=\; \bar{\Lambda}_t X_t \bar{\Lambda}_t' \;+\; \Sigma_{\bar{\omega},t+1} \;=\; X_t \bar{\Lambda}_t'$$

will be of full rank as well. We can then apply (D.8) to obtain

$$I_t \;=\; -\frac{1}{2}\log\frac{\det[(I-\bar{\Lambda}_t)X_t\bar{\Lambda}_t']}{\det[X_t\bar{\Lambda}_t']} \;=\; -\frac{1}{2}\log\det(I-\bar{\Lambda}_t), \qquad \text{(D.10)}$$

in conformity with equation (2.17) in the text.

In the case that X_t is of full rank, but $\bar{\Lambda}_t$ is varied so that one of its eigenvalues approaches 1 (meaning that $I - \bar{\Lambda}_t$ approaches a singular matrix, while the determinant of $\bar{\Lambda}_t$ remains bounded away from zero), the value of I_t implied by (D.10) grows without bound. It thus makes sense to assign a value of $+\infty$ to the mutual information in the case that $\bar{\Lambda}_t$ is of full rank but $I - \bar{\Lambda}_t$ is not. Note that in this case there is a linear combination of the elements of $\bar{s}_t$ that is revealed with perfect precision by the memory state (since $\Sigma_{\bar{\omega},t+1}$ will be singular), while this linear combination is a continuous random variable with positive variance (since X_t is of full rank). This is not consistent with any finite value for the mutual information (and so cannot represent a feasible memory structure).

Suppose instead that while X_t is of full rank, $\bar{\Lambda}_t$ is only of rank one. In this case, we have shown above that $\bar{\Lambda}_t$ must be of the form (D.2), as a consequence of which $\Sigma_{\bar{\omega},t+1}$ must be given by (D.3). In this case, the memory state can be represented in the form $\bar{m}_{t+1} = X_t v_t \cdot \tilde{m}_{t+1}$, where $\tilde{m}_{t+1}$ is a scalar random variable with law of motion (D.4). This implies that $\mathrm{var}[\tilde{m}_{t+1}\,|s_t] \;=\; \mathrm{var}[\tilde{\omega}_{t+1}] \;=\; \lambda_t(1-\lambda_t)$, while $\mathrm{var}[\tilde{m}_{t+1}] \;=\; \lambda_t$. In the case that $0 < \lambda_t < 1$, we can then apply (D.9) to show that

$$I_t \;=\; -\frac{1}{2}\log\frac{\lambda_t(1-\lambda_t)}{\lambda_t} \;=\; -\frac{1}{2}\log(1-\lambda_t), \qquad \text{(D.11)}$$

Since in this case, $\det(I - \hat{\Lambda}_t) = \det(I - \lambda_t v_t v_t') = 1 - \lambda_t$, result (D.11) is again just what (D.10) would imply, so that (D.10) continues to be correct even though $\bar{\Lambda}_t$ is singular.

If we consider a sequence of matrices of this kind in which λ_t approaches 1, the mutual information (D.11) grows without bound. Thus we can assign the value $+\infty$ to I_t in the case that $\bar{\Lambda}_t$ is a matrix of rank one with $\lambda_t = 1$. Indeed, in this case, the memory state reveals with perfect precision the value of $v_t'\bar{s}_t$, a continuous random variable with positive variance (under the assumption that X_t is of full rank); but this is not possible in the case of any finite bound on mutual information. Hence (D.10) can be applied to this case as well.

Suppose instead that X_t is of full rank, but $\bar{\Lambda}_t = 0$. In this case, the distribution of $\bar{m}_{t+1}$ is independent of the value of s_{t+1}, and the mutual information between these two variables must be zero. This is also what (D.10) would imply, so that (D.10) is correct in this case as well.

Finally, consider the case in which $X_t = X_0$, the only possible case in which X_t is not of full rank. In this case, we have defined $\mathcal{L}(X_0)$ to consist only of matrices of the form (D.2), with the vector v_t given by (2.25). If $\lambda_t = 0$, then the entire matrix $\bar{\Lambda}_t = 0$, and the argument in the previous paragraph again applies. Suppose instead that $\lambda_t > 0$. Just as in the discussion above of the case of a singular transition matrix, the memory state can be represented by a scalar state variable $\tilde{m}_{t+1}$ with law of motion (D.4), and we can apply (D.9) to show that I_t will be given by (D.11). Again this is just what (D.10) would imply, so that (D.10) also yields the correct conclusion when X_t is a singular matrix.

Thus in all cases, (D.10) applies, and the value of I_t depends only on the choice of the transition matrix $\bar{\Lambda}_t$. It follows that for any sequence of transition matrices $\{\bar{\Lambda}_t\} \in \mathcal{L}^{seq}$, there will be uniquely defined sequences $\{MSE_t, I_t\}$, allowing the objective (2.5) to be evaluated.

E Recursive Determination of the Optimal Memory Structure

We have shown in the text how the optimal memory structure can be characterized if we can find the value function $V(\hat{\sigma}_t^2)$ that satisfies the Bellman equation

$$V(\hat{\sigma}_t^2) \;=\; \min_{\bar{\Lambda}_t \in \mathcal{L}(X(\hat{\sigma}_t^2))} [\alpha \hat{\sigma}_t^2 \;+\; c(I(\bar{\Lambda}_t)) \;+\; \beta V(f(\hat{\sigma}_t^2, \lambda_t, v_t))]. \tag{E.1}$$

Here we establish some properties of the solution to the optimization problem on the right-hand side of (E.1) for an arbitrary function $V \in \mathcal{F}$., which we can then be used to establish properties of the value function $V(\hat{\sigma}_t^2)$ that solves this equation, and properties of the optimal memory structure.

E.1 Monotonicity of the value function

We first show that, for any function V that may be assumed in the problem on the right-hand side of (E.1), the minimum achievable value of the right-hand side is a monotonically increasing function of $\hat{\sigma}_t^2$. This in turn implies that the value function (which must satisfy (E.1)) must be a monotonically increasing function of its argument.

Fix any value function V to be used in the problem on the right-hand side of (E.1), and consider any two possible degrees of uncertainty $\hat{\sigma}_a^2, \hat{\sigma}_b^2$, satisfying

$$0 \;\leq\; \hat{\sigma}_a^2 \;<\; \hat{\sigma}_b^2 \;\leq\; \sigma_0^2. \tag{E.2}$$

Let $\bar{\Lambda}_t = \bar{\Lambda}_b$ be some element of $\mathcal{L}(X(\hat{\sigma}_b^2))$, and thus a feasible memory structure when $\hat{\sigma}_t^2 = \hat{\sigma}_b^2$, and let us further suppose that $I(\bar{\Lambda}_b) < \infty$, as must be true of an optimal memory structure. We wish to show that we can choose a transition matrix $\bar{\Lambda}_a \in \mathcal{L}(X(\hat{\sigma}_a^2))$ such that

$$f(\hat{\sigma}_a^2, \bar{\Lambda}_a) \;=\; f(\hat{\sigma}_b^2, \bar{\Lambda}_b), \tag{E.3}$$

and in addition

$$I(\bar{\Lambda}_a) \;\leq\; I(\bar{\Lambda}_b). \tag{E.4}$$

That is, in the case of the smaller degree of uncertainty $\hat{\sigma}_a^2$ in the cognitive state in period t, it is possible to choose a memory structure that implies exactly the same degree of uncertainty in period $t+1$, and hence the same value for $V(\hat{\sigma}_{t+1}^2)$, at no greater an information cost, and thus it is possible to achieve a strictly lower value for the right-hand side of (E.1).

If we can show this for an arbitrary transition matrix $\bar{\Lambda}_b \in \mathcal{L}(X(\hat{\sigma}_b^2))$, then it is also true when $\bar{\Lambda}_b$ is the transition matrix associated with the optimal memory structure (the solution to the problem on the right-hand side of (E.1)) when $\hat{\sigma}_t^2 = \hat{\sigma}_b^2$. This implies that it is possible to achieve a lower value for the right-hand side of (E.1) when $\hat{\sigma}_t^2 = \hat{\sigma}_a^2$ than it is possible to achieve when $\hat{\sigma}_t^2 = \hat{\sigma}_b^2$. Since this must be true for any values of $\hat{\sigma}_a^2, \hat{\sigma}_b^2$ consistent with (E.2), the right-hand side of (E.1) defines a monotonically increasing function of $\hat{\sigma}_t^2$.

To show that such a construction is always possible, let us first consider the case in which $\hat{\sigma}_b^2 = \hat{\sigma}_0^2$, so that the memory state m_t is completely uninformative in case b. In this case, the assumption that $\bar{\Lambda}_b \in \mathcal{L}(X(\hat{\sigma}_b^2)) = \mathcal{L}(X_0)$ requires that

$$\bar{\Lambda}_b \;=\; \lambda_b \frac{ww'}{w'w}$$

for some $0 \leq \lambda_b < 1$.[72] In this case, the memory structure for the following period is equivalent to one in which there is a univariate memory state

$$\tilde{m}_b \;=\; \frac{\lambda_b}{(\Omega + \sigma_y^2)^{1/2}} y_t \,+\, \tilde{\omega}_b, \qquad \tilde{\omega}_b \,\sim\, N(0,\, \lambda_b(1-\lambda_b)).$$

The implied uncertainty in the following period (given the memory state, but before y_{t+1} is observed) is then given by

$$\Sigma_{t+1} \;=\; \Sigma_0 \,-\, \lambda_b(\Omega + \sigma_y^2)ww'. \tag{E.5}$$

Now let $\bar{s}_a$ be the reduced cognitive state in period t, in the case of a more informative memory structure that implies the lower degree of uncertainty $\hat{\sigma}_a^2$, and let $X_a \equiv X(\hat{\sigma}_a^2)$ be the variance of this random vector. In this case, we can choose a memory structure for the following period defined by the transition matrix

$$\bar{\Lambda}_a \;=\; \lambda_b X_a \frac{e_2 e_2'}{\Omega + \sigma_y^2}$$

where $e_2 \equiv [0\ 1]'$. This is a matrix of the form (D.2), and hence an element of $\mathcal{L}(X_a)$. Because $\bar{\Lambda}_a$ is singular, the specified memory structure is equivalent to one in which there is a univariate memory state

$$\tilde{m}_a \;=\; \lambda_b \frac{e_2'\bar{s}_a}{(e_2' X_a e_2)^{1/2}} \,+\, \tilde{\omega}_a, \qquad \tilde{\omega}_a \,\sim\, N(0,\, \lambda_b(1-\lambda_b)).$$

But this means that

$$\tilde{m}_a \;=\; \frac{\lambda_b}{(\Omega + \sigma_y^2)^{1/2}} y_t \,+\, \tilde{\omega}_a, \qquad \tilde{\omega}_a \,\sim\, N(0,\, \lambda_b(1-\lambda_b)).$$

Hence the joint distribution of $(\tilde{m}_a,\, x_{t+1})$ is identical to the joint distribution of $(\tilde{m}_b,\, x_{t+1})$, and the implied uncertainty in the following period given this memory structure is again given by (E.5). Hence the value of $\hat{\sigma}_{t+1}^2$ implied by memory structure a is the same as that

[72] The upper bound is required in order to satisfy the assumption that $I(\bar{\Lambda}_b) < \infty$.

implied by memory structure b. This establishes condition (E.3). Moreover, for both memory structures we have the same mutual information,

$$I(\bar{\Lambda}_a) \;=\; I(\bar{\Lambda}_b) \;=\; -\frac{1}{2}\log(1-\lambda_b).$$

This establishes condition (E.4). Hence the value of the right-hand side of (E.1) must be lower when $\hat{\sigma}_t^2 = \hat{\sigma}_a^2$.

Let us next consider the less trivial case in which $0 < \hat{\sigma}_b^2 < \hat{\sigma}_0^2$. Let $\bar{s}_b$ be the reduced cognitive state in period t that implies a degree of uncertainty $\hat{\sigma}_b^2$, and let $X_b \equiv X(\hat{\sigma}_b^2)$ be the variance of this random vector. Let the optimal memory structure for the following period (the solution to the problem on the right-hand side of (E.1)) in this case be

$$\bar{m}_b \;=\; \bar{\Lambda}_b\bar{s}_b \,+\, \bar{\omega}_b, \tag{E.6}$$

where

$$\bar{\Lambda}_b \,\in\, \mathcal{L}(X_b), \qquad \bar{\omega}_b \,\sim\, N(0,\,(I-\bar{\Lambda}_b)X_b\bar{\Lambda}_b').$$

The implied uncertainty in the following period will then be given by

$$\Sigma_{t+1} \;=\; \Sigma_0 \,-\, X_b\bar{\Lambda}_b'. \tag{E.7}$$

Let us consider the memory structure for cognitive state a defined by the transition matrix

$$\bar{\Lambda}_a \;=\; \bar{\Lambda}_b\Gamma\Psi\Gamma^{-1}, \tag{E.8}$$

where Γ is the invertible matrix defined in (C.5), and

$$\Psi \;\equiv\; \begin{bmatrix} \psi & 0 \\ 0 & 1 \end{bmatrix},$$

where $0 < \psi < 1$ is the quantity

$$\psi \;\equiv\; \frac{\hat{\sigma}_0^2 - \hat{\sigma}_b^2}{\hat{\sigma}_0^2 - \hat{\sigma}_a^2}.$$

Note that Ψ is a diagonal matrix, with the property that

$$\Psi\check{X}_a \;=\; \check{X}_a\Psi \;=\; \check{X}_b,$$

using the notation $\check{X}_i \equiv \check{X}(\hat{\sigma}_i^2)$ for $i = a, b$, where $\check{X}(\hat{\sigma}_t^2)$ is the function defined in (C.6). It is first necessary to verify that $\bar{\Lambda}_a \in \mathcal{L}(X_a)$, so that this matrix defines a possible memory structure.

We first show that $\bar{\Lambda}_a X_a = X_a\bar{\Lambda}_a'$. Definition (E.8) implies that

$$\begin{aligned} \bar{\Lambda}_a X_a &= \bar{\Lambda}_b\Gamma\Psi\Gamma^{-1}X_a \\ &= \bar{\Lambda}_b\Gamma\Psi\check{X}_a\Gamma' \\ &= \bar{\Lambda}_b\Gamma\check{X}_b\Gamma' \\ &= \bar{\Lambda}_b X_b. \end{aligned}$$

The fact that $\bar{\Lambda}_b \in \mathcal{L}(X_b)$ implies that $\bar{\Lambda}_b X_b$ must be a symmetric matrix; hence $\bar{\Lambda}_a X_a$, which is the same matrix, must also be symmetric. Thus $\bar{\Lambda}_a X_a = X_a \bar{\Lambda}_a'$.

Next, we must also show that $(I - \bar{\Lambda}_a) X_a \bar{\Lambda}_a'$ is a p.s.d. matrix. We first note that $I - \Psi$ is a diagonal matrix with non-negative elements on the diagonal; it follows that $(I - \Psi)\check{X}_b$ is also a diagonal matrix with non-negative elements on the diagonal, and hence p.s.d. From this it follows that

$$\begin{aligned}
\bar{\Lambda}_b \Gamma \cdot (I - \Psi)\check{X}_b \cdot \Gamma' \bar{\Lambda}_b' &= \bar{\Lambda}_b \Gamma (\check{X}_b - \Psi \check{X}_a \Psi) \Gamma' \bar{\Lambda}_b' \\
&= \bar{\Lambda}_b (\Gamma \check{X}_b \Gamma') \bar{\Lambda}_b' \; - \; (\bar{\Lambda}_b \Gamma \Psi \Gamma^{-1})(\Gamma \check{X}_a \Gamma')(\bar{\Lambda}_b \Gamma \Psi \Gamma^{-1})' \\
&= \bar{\Lambda}_b X_b \bar{\Lambda}_b' \; - \; \bar{\Lambda}_a X_a \bar{\Lambda}_a' \\
&= (X_a \bar{\Lambda}_a' \; - \; \bar{\Lambda}_a X_a \bar{\Lambda}_a') \; - \; (X_b \bar{\Lambda}_b' \; - \; \bar{\Lambda}_b X_b \bar{\Lambda}_b') \\
&= (I - \bar{\Lambda}_a) X_a \bar{\Lambda}_a' \; - \; (I - \bar{\Lambda}_b) X_b \bar{\Lambda}_b'
\end{aligned}$$

must be p.s.d. as well. But since the fact that $\bar{\Lambda}_b \in \mathcal{L}(X_b)$ implies that $(I - \bar{\Lambda}_b) X_b \bar{\Lambda}_b'$ must be p.s.d., it follows that $(I - \bar{\Lambda}_a) X_a \bar{\Lambda}_a'$ can be expressed as the sum of two p.s.d. matrices, and so must also be p.s.d. This verifies the second of the conditions required in order to show that $\bar{\Lambda}_a \in \mathcal{L}(X_a)$.

Thus if $\bar{s}_a$ is a reduced cognitive state for period t that implies a degree of uncertainty $\hat{\sigma}_a^2$, a possible memory structure for the following period is

$$\bar{m}_a \; = \; \bar{\Lambda}_a \bar{s}_a \; + \; \bar{\omega}_a, \tag{E.9}$$

where the transition matrix $\bar{\Lambda}_a$ is defined in (E.8), and

$$\bar{\omega}_a \; \sim \; N(0, \, (I - \bar{\Lambda}_a) X_a \bar{\Lambda}_a').$$

The implied uncertainty in the following period will then be given by

$$\Sigma_{t+1} \; = \; \Sigma_0 \; - \; X_a \bar{\Lambda}_a'.$$

This latter matrix is the same as the one in (E.7); it follows that the implied value of $\hat{\sigma}_{t+1}^2$ is also the same as for the memory structure (E.6). Thus we have shown that in the case of the smaller degree of uncertainty $\hat{\sigma}_a^2$, it is possible to choose a memory structure that implies exactly the same degree of uncertainty in period $t+1$ as when the degree of uncertainty in period t is given by the larger quantity $\hat{\sigma}_b^2$.

It remains to be shown that memory structure (E.9) involves no greater information cost than memory structure (E.6). Consider first the case in which the memory state $\bar{m}_b$ is non-degenerate, in the sense that $\text{var}[\bar{m}_b] = X_b \bar{\Lambda}_b'$ is non-singular. It follows that the same must be true of memory state $\bar{m}_a$. Then for either of the two memory structures $i = a, b$ just discussed, (D.10) implies that the mutual information will be given by

$$I_t \; = \; -\frac{1}{2} \log \frac{\det[(I - \bar{\Lambda}_i) X_i \bar{\Lambda}_i']}{\det[X_i \bar{\Lambda}_i']}.$$

We have shown above that the value of the denominator in this expression is the same for $i = a, b$ (and under the assumption that $X_b \bar{\Lambda}_b'$ is non-singular, it must be positive). Hence the relative size of the two mutual informations depends on the relative size of the numerator in the two cases. But we have shown above that $(I - \bar{\Lambda}_a)X_a\bar{\Lambda}'_a$ can be expressed as the sum of $(I - \bar{\Lambda}_b)X_b\bar{\Lambda}'_b$ plus a p.s.d. matrix. Since both of these matrices are also p.s.d., their determinants satisfy

$$\det[(I - \bar{\Lambda}_a)X_a\bar{\Lambda}'_a] \;\geq\; \det[(I - \bar{\Lambda}_b)X_b\bar{\Lambda}'_b] \;>\; 0,$$

where the final inequality is necessary in order for memory structure b to have a finite information cost. It follows that condition (E.4) must hold in this case.

Now suppose instead that $\text{var}[\bar{m}_b]$ is a singular matrix. In the case that the matrix is zero in all elements, $\bar{\Lambda}_b = 0$, and so (E.8) implies that $\bar{\Lambda}_a = 0$ as well. In this case, $\det(I - \bar{\Lambda}_a) = \det(I - \bar{\Lambda}_b) = 1$, so that $I(\bar{\Lambda}_a) = I(\bar{\Lambda}_b) = 0$, and (E.4) is satisfied in this case as well. Thus we need only consider further the case in which $\text{var}[\bar{m}_b]$ is of rank one, which requires that $\bar{\Lambda}_b$ be of rank one as well.

In this case, we can write

$$\bar{\Lambda}_b \;=\; \lambda_b X_b v_b v'_b,$$

where $0 < \lambda_b < 1$[73] and v_b is a vector such that $v'_b X_b v_b = 1$. All columns of $\bar{\Lambda}_b$ are multiples of the vector $X_b v_b$, and as a consequence the unique non-null right eigenvector of $\bar{\Lambda}_b$ is given by $X_b v_b$, with the associated eigenvalue λ_b. Alternatively, using the orthogonalized representation of the cognitive state introduced in section C.4, we can write

$$\Gamma^{-1}\bar{\Lambda}_b\Gamma \;=\; \lambda_b \check{X}_b \check{v}_b \check{v}'_b,$$

where we define $\check{v}_b \equiv \Gamma' v_b$, and note that $\check{v}'_b \check{X}_b \check{v}_b = 1$.

Then (E.8) implies that the columns of $\bar{\Lambda}_a$ must also all be multiples of the vector $X_b v_b$. It follows that $\bar{\Lambda}_a$ must also be singular, and that its unique non-null eigenvector must be $X_b v_b$, with an associated eigenvalue

$$\begin{aligned}
\lambda_a \;&=\; \lambda_b v'_b \Gamma\Psi\Gamma^{-1}(X_b v_b) \\
&=\; \lambda_b \check{v}'_b \Psi \check{X}_b \check{v}_b \\
&=\; \lambda_b (\check{v}'_b \Psi^{1/2}) \check{X}_b (\Psi^{1/2} \check{v}_b) \\
&\leq\; \lambda_b \check{v}'_b \check{X}_b \check{v}_b \;=\; \lambda_b.
\end{aligned}$$

Thus we must have

$$\det(I - \bar{\Lambda}_a) \;=\; (1 - \lambda_a) \;\geq\; (1 - \lambda_b) \;=\; \det(I - \bar{\Lambda}_b),$$

from which it follows that (E.4) must hold in this case as well.

Thus we have shown that whenever $\hat{\sigma}^2_a, \hat{\sigma}^2_b$ satisfy (E.2), for any memory structure for case b with a finite information cost, it is possible to choose a memory stucture for case a satisfying both (E.3) and (E.4). This means that it must be possible to achieve a lower value for the right-hand side of (E.1) when $\hat{\sigma}^2_t = \hat{\sigma}^2_a$ than when $\hat{\sigma}^2_b$. This in turn implies that the right-hand side of (E.1) defines a monotonically increasing function of $\hat{\sigma}^2_t$, regardless of the nature of the function $V(\hat{\sigma}^2_{t+1})$ that is assumed in this optimization problem. Hence the value function $V(\hat{\sigma}^2_t)$ defined by (E.1) must be a monotonically increasing function of its argument.

[73] Again, the upper bound is required in order for $I(\bar{\Lambda}_b)$ to be finite.

E.2 Optimality of a unidimensional memory state

Here we establish, as stated in the text, that the matrix $\bar{\Lambda}_t$ that solves the problem

$$\min_{\bar{\Lambda}_t \in \mathcal{L}(X(\hat{\sigma}_t^2))} I(\bar{\Lambda}_t) \qquad \text{s.t. } f(\hat{\sigma}_t^2, \bar{\Lambda}_t) \leq \hat{\sigma}_{t+1}^2, \tag{E.10}$$

for given values of $(\hat{\sigma}_t^2, \hat{\sigma}_{t+1}^2)$ is necessarily at most of rank one. As explained in the text, we need only consider the case in which $\hat{\sigma}_t^2 < \hat{\sigma}_0^2$. Given a matrix $\bar{\Lambda}_t$ of rank two that satisfies the constraint in (E.10), we wish to show that we can choose an alternative transition matrix of at most rank one, that also satisfies the constraint, but which achieves a lower value of $I(\bar{\Lambda}_t)$.

We first note that when $\hat{\sigma}_t^2 < \hat{\sigma}_0^2$, $X(\hat{\sigma}_t^2)$ is non-singular. Under the hypothesis that $\bar{\Lambda}_t$ is non-singular, it follows that $X_t\bar{\Lambda}_t'$ is non-singular as well (where we now simply write X_t for $X(\hat{\sigma}_t^2)$), and hence positive definite. Similarly, $\bar{\Lambda}_t X_t \bar{\Lambda}_t'$ must be non-singular and hence positive definite.

Then let the alternative transition matrix be given by

$$\bar{\Lambda}_t^{1D} = \lambda_t X_t v_t v_t', \tag{E.11}$$

with

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

where $\delta_{t+1} \equiv e_1 - \gamma_{1,t+1}c$ is the vector introduced in (2.21), and let the matrix $\Sigma_{\bar{\omega},t+1}$ be correspondingly modified in the way specified by (2.14). The fact that $X_t\bar{\Lambda}_t'$ is positive definite implies that the denominator of the expression for λ_t is necessarily positive, so that this quantity is well-defined. Similarly, the fact that $\bar{\Lambda}_t X_t \bar{\Lambda}_t'$ is positive definite implies that the denominator of the expression for v_t is necessarily positive, so that this vector is well-defined as well.

In addition, the fact that (by assumption) $\bar{\Lambda}_t \in \mathcal{L}(X_t)$ implies that $(I - \bar{\Lambda}_t)X_t\bar{\Lambda}_t'$ must be p.s.d. From this it follows that

$$\delta_{t+1}'(I - \bar{\Lambda}_t)X_t\bar{\Lambda}_t'\delta_{t+1} \geq 0,$$

and hence that

$$\delta_{t+1}'X_t\bar{\Lambda}_t'\delta_{t+1} \geq \delta_{t+1}'\bar{\Lambda}_t X_t\bar{\Lambda}_t'\delta_{t+1} > 0,$$

where the final inequality follows from the fact that $\bar{\Lambda}_t X_t \bar{\Lambda}_t'$ is positive definite. Thus the proposed definition of λ_t satisfies $0 < \lambda_t \leq 1$. One also observes from the definition of v_t that $v_t'X_tv_t = 1$. These conditions suffice to establish that the alternative transition matrix $\bar{\Lambda}_t^{1D}$ is also an element of $\mathcal{L}(X_t)$. That is, it represents a feasible memory structure for period t, given the value of $\hat{\sigma}_t^2$.

This alternative transition matrix corresponds to a memory structure in which $\bar{m}_{t+1} = X_t v_t \tilde{m}_{t+1}$, where $\tilde{m}_{t+1}$ is the unidimensional memory state with law of motion (2.24). From this it follows that

$$\delta_{t+1}'\bar{m}_{t+1} = \lambda_t \delta_{t+1}' X_t v_t v_t' \bar{s}_t + \delta_{t+1}' X_t v_t \tilde{\omega}_{t+1}$$

will be a normally distributed random variable, with conditional first and second moments given by

$$\begin{aligned}
\mathrm{E}[\delta'_{t+1}\bar{m}_{t+1}\,|s_t] &= \lambda_t\delta'_{t+1}X_t v_t v'_t\bar{s}_t \\
&= \frac{\delta'_{t+1}\bar{\Lambda}_t X_t\bar{\Lambda}'_t\delta_{t+1}}{\delta'_{t+1}X_t\bar{\Lambda}'_t\delta_{t+1}}\frac{\delta'_{t+1}X_t\bar{\Lambda}'_t\delta_{t+1}\cdot\delta'_{t+1}\bar{\Lambda}_t\bar{s}_t}{\delta'_{t+1}\bar{\Lambda}_t X_t\bar{\Lambda}'_t\delta_{t+1}} \\
&= \delta'_{t+1}\bar{\Lambda}_t\bar{s}_t
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{var}[\delta'_{t+1}\bar{m}_{t+1}\,|s_t] &= \lambda_t(1-\lambda_t)(\delta'_{t+1}X_t v_t)^2 \\
&= (1-\lambda_t)\frac{\delta'_{t+1}\bar{\Lambda}_t X_t\bar{\Lambda}'_t\delta_{t+1}}{\delta'_{t+1}X_t\bar{\Lambda}'_t\delta_{t+1}}\frac{(\delta'_{t+1}X_t\bar{\Lambda}'_t\delta_{t+1})^2}{\delta'_{t+1}\bar{\Lambda}_t X_t\bar{\Lambda}'_t\delta_{t+1}} \\
&= (1-\lambda_t)\delta'_{t+1}X_t\bar{\Lambda}'_t\delta_{t+1} \\
&= \delta'_{t+1}X_t\bar{\Lambda}'_t\delta_{t+1} - \delta'_{t+1}\bar{\Lambda}_t X_t\bar{\Lambda}'_t\delta_{t+1} \\
&= \delta'_{t+1}[(I-\bar{\Lambda}_t)X_t\bar{\Lambda}'_t]\delta_{t+1} \\
&= \delta'_{t+1}\Sigma_{\bar{\omega}_{t+1}}\delta_{t+1}.
\end{aligned}$$

These are the same conditional mean and variance as in the case of the memory structure specified by the transition matrix $\bar{\Lambda}_t$. Since the optimal estimate $\hat{\mu}_{t+1}$ depends on m_{t+1} only through the value of $\delta'_{t+1}\bar{m}_{t+1}$ (from equation (2.21)), it follows that the conditional distribution $\hat{\mu}_{t+1}|s_t, y_{t+1}$ will be the same under the alternative memory structure. This in turn implies that the variance of $\hat{\mu}_{t+1}$ will be the same, and hence that

$$\hat{\sigma}^2_{t+1} = \Omega - \mathrm{var}[\hat{\mu}_{t+1}]$$

will be the same. Thus $\bar{\Lambda}^{1D}_t$ also satisfies the constraint in (E.10).

Next we show that $I(\bar{\Lambda}^{1D}_t)$ must be lower than $I(\bar{\Lambda}_t)$. Let u'_1 and u'_2 be the two left eigenvectors of $\bar{\Lambda}_t$, with associated eigenvalues μ_1 and μ_2 respectively, and let the eigenvectors be normalized so that $u'_i X_t u_i = 1$ for $i = 1, 2$. The corresponding right eigenvectors must then be $X_t u_1$ and $X_t u_2$ respectively. Thus we have

$$\bar{\Lambda}_t X_t u_i = \mu_i X_t u_i, \qquad u'_i\bar{\Lambda}_t = \mu_i u'_i,$$

for $i = 1, 2$, and

$$u'_1 X_t u_1 = u'_2 X_t u_2 = 1, \qquad u'_1 X_t u_2 = 0.$$

The vector δ'_{t+1} introduced in (2.21) can be written as a linear combination of the two left eigenvectors,

$$\delta'_{t+1} = \alpha_1 u'_1 + \alpha_2 u'_2,$$

for some coefficients α_1, α_2. This implies that

$$\delta'_{t+1}X_t\bar{\Lambda}'_t\delta_{t+1} = \alpha_1^2\mu_1 + \alpha_2^2\mu_2,$$

$$\delta'_{t+1}\bar{\Lambda}_t X_t\bar{\Lambda}'_t\delta_{t+1} = \alpha_1^2\mu_1^2 + \alpha_2^2\mu_2^2,$$

and hence that

$$\lambda_t \;=\; \frac{\alpha_1^2\mu_1}{\alpha_1^2\mu_1 + \alpha_2^2\mu_2}\mu_1 \;+\; \frac{\alpha_2^2\mu_2}{\alpha_1^2\mu_1 + \alpha_2^2\mu_2}\mu_2.$$

Thus we see that λ_t must be a convex combination of μ_1 and μ_2.

The fact that $\bar{\Lambda}_t \in \mathcal{L}(X_t)$ requires that both eigenvalues satisfy $0 \leq \mu_i \leq 1$, and the assumption that $\bar{\Lambda}_t$ is non-singular further requires that $\mu_i > 0$ for both. Thus we must have

$$1 - \mu_i \;> (1 - \mu_1)(1 - \mu_2)$$

for both $i = 1, 2$. Since λ_t is a convex combination of μ_1 and μ_2, it follows that

$$1 - \lambda_t \;>\; (1 - \mu_1)(1 - \mu_2).$$

Thus

$$\det(I - \bar{\Lambda}_t^{1D}) \;=\; 1 - \lambda_t \;>\; (1 - \mu_1)(1 - \mu_2) \;=\; \det(I - \bar{\Lambda}_t).$$

Results (D.10) and (D.11) then imply that $I(\bar{\Lambda}_t^{1D}) < I(\bar{\Lambda}_t)$.

Thus $\bar{\Lambda}_t$ cannot be the solution to the optimization problem (E.10). Since this argument can be made in the case of any matrix $\bar{\Lambda}_t \in \mathcal{L}(X_t)$ that is of full rank, we conclude that the optimal transition matrix can be at most of rank one.

E.3 The optimal univariate memory state

We turn now to the question of which linear combination of the elements of the reduced cognitive state constitutes the single variable for which it is optimal to retain a noisy record in memory — that is, we wish to characterize the optimal weight vector v_t in (D.4). Here we take as given the value of λ_t (or equivalently, the mutual information between the period t cognitive state and the memory carried into period $t+1$), and solve for the optimal choice of v_t for any given value of λ_t. With this in hand, it will then be possible to characterize an optimal memory structure in terms of the single parameter λ_t.

Given the value of $\hat{\sigma}_t^2$ and the matrix $X_t \equiv \mathrm{var}[\bar{s}_t]$, and taking as given the value of λ_t, we wish to choose v_t so as to minimize $\hat{\sigma}_{t+1}^2$. Note that

$$\hat{\sigma}_{t+1}^2 \;=\; \min_{\xi,\gamma_1} \mathrm{var}[\mu - \xi\tilde{m}_{t+1} - \gamma_1 y_{t+1}].$$

Hence we can write our problem as the choice of ξ, γ_1, and the vector v_t so as to minimize

$$\begin{aligned} f(\hat{\sigma}_t^2, \lambda_t, v_t;\, \xi, \gamma_1) \;&\equiv\; \mathrm{var}[\mu \;-\; \xi(\lambda_t v_t'\bar{s}_t + \tilde{\omega}_{t+1}) \;-\; \gamma_1 y_{t+1}] \\ &=\; \mathrm{var}[\mu \;-\; \xi\lambda_t v_t'\bar{s}_t \;-\; \gamma_1 y_{t+1}] \;+\; \xi^2\lambda_t(1 - \lambda_t), \end{aligned}$$

subject to the constraint that $v_t'X_t v_t = 1$. Note that the solution to this problem will simultaneously determine the optimal choice of v_t (and hence the optimal memory structure, given a choice of λ_t) and the coefficients of the optimal estimate

$$\hat{\mu}_{t+1} \;=\; \xi\tilde{m}_{t+1} \;+\; \gamma_1 y_{t+1} \tag{E.12}$$

based on that memory structure.

We can alternatively define this problem as the choice of a weighting vector $\psi \equiv \xi \lambda_t v_t$ and a Kalman gain γ_1. The values of these quantities suffice to determine the value of the objective (if we know the values of $\hat{\sigma}_t^2$ and λ_t), since we can reconstruct ξ and v_t from them:

$$v_t \;=\; \frac{\psi}{(\psi' X_t \psi)^{1/2}}, \qquad \xi \;=\; (\psi' X_t \psi)^{1/2} \lambda_t.$$

Moreover, there is no theoretical restriction on the elements of the vector ψ, since the scale factor ξ can be of arbitrary size in the previous formulation of the optimization problem. Thus we can alternatively state our problem as the choice of a weighting vector ψ and a Kalman gain γ_1 to minimize

$$f(\hat{\sigma}_t^2, \lambda_t;\, \psi, \gamma_1) \;=\; \mathrm{var}[\mu \,-\, \psi' \bar{s}_t \,-\, \gamma_1 y_{t+1}] + \frac{1-\lambda_t}{\lambda_t} \psi' X_t \psi. \tag{E.13}$$

We can write the first term in this objective as

$$\begin{aligned} \mathrm{var}[\mu \,-\, \psi' \bar{s}_t \,-\, \gamma_1 y_{t+1}] \;&=\; \mathrm{var}[(1-(1-\rho)\gamma_1)(\mu - \hat{\mu}_t) - \gamma_1 (y_{t+1} - \mu) + (e_1' - \gamma_1 c')\bar{s}_t - \psi' \bar{s}_t] \\ &=\; (e_1' - \gamma_1 c' - \psi') X_t (e_1 - \gamma_1 c - \psi) \,+\, (1-(1-\rho)\gamma_1)^2 \hat{\sigma}_t^2 \,+\, \gamma_1^2 \sigma_\epsilon^2. \end{aligned}$$

Substituting this into (E.13), we see that the objective is a strictly convex quadratic function of ψ and γ_1, for any values of $\hat{\sigma}_t^2$ and λ_t. It follows that the objective has an interior minimum, given by the unique solution to the first-order conditions.

The FOCs for the minimization of (E.13) are given by the linear equations

$$\psi \;=\; \lambda_t (e_1 - \gamma_1 c), \tag{E.14}$$

$$c' X_t (e_1 - \gamma_1 c - \psi) \,+\, (1-\rho)(1-(1-\rho)\gamma_1)\hat{\sigma}_t^2 \,-\, \gamma_1 \sigma_\epsilon^2 \;=\; 0. \tag{E.15}$$

Equation (E.14) already allows one valuable insight: the optimal weight vector v_t is simply a normalized version of the vector δ_{t+1} defined in (2.21). However, this does not yet tell us how to choose v_t, since the vector δ_{t+1} depends on the Kalman gain $\gamma_{1,t+1}$, which depends on the memory structure chosen in period t.

But together equations (E.14)–(E.15) provide a linear system that can be solved for ψ and γ_1, given the values of $\hat{\sigma}_t^2$ and λ_t. We obtain

$$\gamma_{1,t+1} \;=\; \frac{(1-\lambda_t)\Omega + \lambda_t (1-\rho)\hat{\sigma}_t^2}{(1-\lambda_t)(\Omega + \rho^2 \sigma_y^2) + \lambda_t (1-\rho)^2 \hat{\sigma}_t^2 + \sigma_\epsilon^2} \tag{E.16}$$

as an explicit solution for the Kalman gain. It is worth noting that this implies that

$$0 \;<\; \gamma_{1,t+1} \;<\; \frac{1}{1-\rho}. \tag{E.17}$$

We can then use this solution to evaluate the elements of the vector δ. We obtain

$$\delta_{1,t+1} \;\equiv\; 1 - (1-\rho)\gamma_{1,t+1} \;=\; \frac{(1-\lambda_t)\rho(\Omega + \rho\sigma_y^2) + \sigma_\epsilon^2}{(1-\lambda_t)(\Omega + \rho^2 \sigma_y^2) + \lambda_t (1-\rho)^2 \hat{\sigma}_t^2 + \sigma_\epsilon^2} \;>\; 0,$$

$$\delta_{2,t+1} \;\equiv\; -\rho\gamma_{1,t+1} \;=\; -\frac{(1-\lambda_t)\rho\Omega + \lambda_t\rho(1-\rho)\hat{\sigma}_t^2}{(1-\lambda_t)(\Omega + \rho^2\sigma_y^2) + \lambda_t(1-\rho)^2\hat{\sigma}_t^2 + \sigma_\epsilon^2} \;\leq\; 0.$$

The weight vector v_t is then just a normalized version of δ_{t+1}.

We note that when $\rho = 0$, the optimal weight vector has $v_2 = 0$; that is, the memory state $\tilde{m}_{t+1}$ is just a noisy record of $\hat{\mu}_t$. (This is intuitive, since when the state is i.i.d., and given the estimate $\hat{\mu}_t$ of the mean, the value of y_t provides no information about anything that needs to be estimated or forecasted in period $t+1$ or later.) Instead when $\rho > 0$, we see that the sign of v_2 is necessarily opposite to the sign of v_1: the optimal memory state averages $\hat{\mu}_t$ and y_t with a negative relative weight on y_t.

Given this solution for γ_1, the implied solution for the vector ψ is given by (E.14). Substituting the solutions for γ_1 and ψ into the quadratic objective, we obtain for the minimum possible value of the objective

$$\hat{\sigma}_{t+1}^2 \;=\; (1-\lambda_t)\delta_{t+1}'\Sigma_0\delta_{t+1} + \lambda_t(\delta_{1,t+1})^2\hat{\sigma}_t^2 \;+\; \gamma_{1,t+1}^2\sigma_\epsilon^2. \tag{E.18}$$

This provides an equation for the evolution of the uncertainty measure $\hat{\sigma}_{t+1}^2$, given a choice each period of λ_t, and using the formulas above for the values of $\gamma_{1,t+1}$ and δ_{t+1}.

F Numerical Solutions

Here we provide further details of the numerical calculations reported in section 3 of the main text.

F.1 Dynamics of uncertainty given the path of $\{\lambda_t\}$

We begin by discussing our approach to numerical solution for the law of motion $\eta_{t+1} = \phi(\eta_t; \lambda_t)$ for the scaled uncertainty measure $\{\eta_t\}$, given a path for the memory-sensitivity coefficient $\{\lambda_t\}$. In terms of this rescaled state variable, the law of motion (E.18) becomes

$$\eta_{t+1} \;=\; (1-\lambda_t)(1-\gamma_{1,t+1})^2 K \;+\; (1-\rho^2\lambda_t)\gamma_{1,t+1}^2 \;+\; \lambda_t(1-(1-\rho)\gamma_{1,t+1})^2\eta_t, \tag{F.1}$$

and (E.16) becomes

$$\gamma_{1,t+1} \;=\; \frac{(1-\lambda_t)K + (1-\rho)\lambda_t\eta_t}{(1-\lambda_t)(K+\rho^2) + (1-\rho^2) + (1-\rho)^2\lambda_t\eta_t}. \tag{F.2}$$

Substitution of (F.2) for $\gamma_{1,t+1}$ in the right-hand side of (F.1) yields an analytical expression for the function $\phi(\eta_t; \lambda_t)$.

This result suffices to allow us to compute the optimal dynamics of the uncertainty measure $\{\eta_t\}$ in the case that the only limit on the complexity of memory is an upper bound $\lambda_t \leq \bar{\lambda} < 1$ each period. We observe from (E.13) that the objective $f(\hat{\sigma}_t^2, \lambda_t; \psi, \gamma_1)$ is minimized, for given values of the other parameters, by making λ_t as large as possible. Hence the same is true for the function $f(\hat{\sigma}_t^2, \lambda_t, v_t)$ obtained by minimizing the objective over possible choices of ξ and γ_1. It follows that it will be optimal to choose $\lambda_t = \bar{\lambda}$ each period in the case of this kind of constraint.

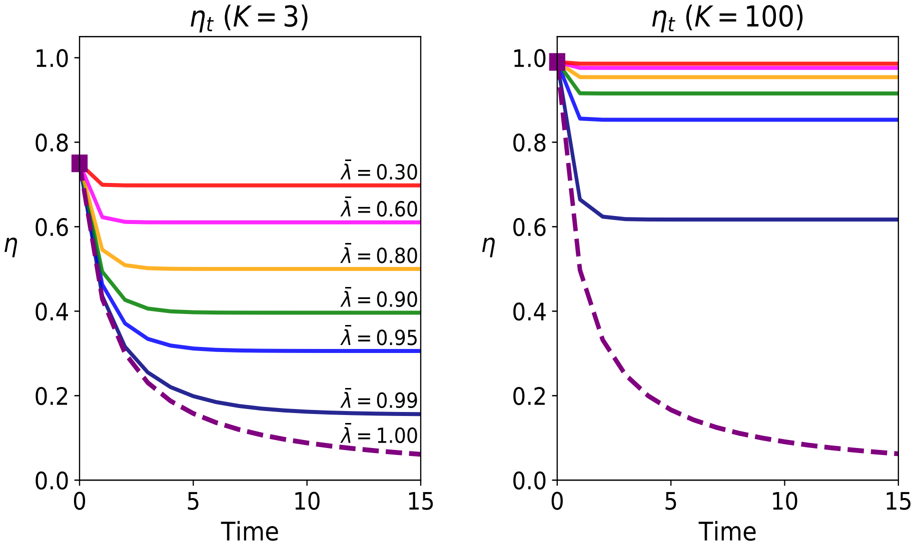


Figure 8: The evolution of scaled uncertainty about μ as the number t of previous (imperfectly remembered) observations grows. Each panel corresponds to a particular value of K (maintaining the assumption that $\rho = 0$, as in Figure 1). Each panel shows the evolution for several different possible values of $\bar{\lambda}$ (color code is the same in both panels).

We thus obtain a nonlinear difference equation

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

for the dynamics of the scaled uncertainty measure. We can iterate this mapping, starting from the initial condition $\eta_0 = K/(K+1)$, to obtain the complete sequence of values $\{\eta_t\}$ for all $t \geq 0$ implied by any given value of $\bar{\lambda}$. This is the method used to compute the dynamic paths shown in Figure 1 in the main text.

Figure 1 shows the dynamics for $\{\eta_t\}$ implied by this solution, for various possible values of $\bar{\lambda}$, in the case that $K = 1$ and $\rho = 0$. Figure 8 shows how this graph would be different in the case of two larger values for K (but again assuming $\rho = 0$). A higher value of K (greater prior uncertainty) implies a higher value for the initial value η_0 of our normalized measure of uncertainty (since $\eta_0 = K/(K+1)$). This means that the curves all start higher, the larger the value of K. But the value of K also affects the long-run level of uncertainty η_∞, even though the initial condition becomes irrelevant in the long run. Except when $\bar{\lambda} = 1$ (perfect memory), a higher value of K implies greater long-run uncertainty; and when K is large (as illustrated in the right panel), η_∞ is large (not much below the degree of uncertainty implied by the prior) except in the case of quite high values of $\bar{\lambda}$.

Figure 9 similarly shows how Figure 1 would look in the case of two larger values of ρ, but again assuming $K = 1$. We see that for a given degree of prior uncertainty and a given bound on memory precision, the rate at which uncertainty is reduced is slower when the external state is more serially correlated. This is because there are effectively fewer independent observations over a given number of periods when the state is serially correlated. In the case of perfect memory ($\bar{\lambda} = 1$), this affects the speed of learning but not the long-run value

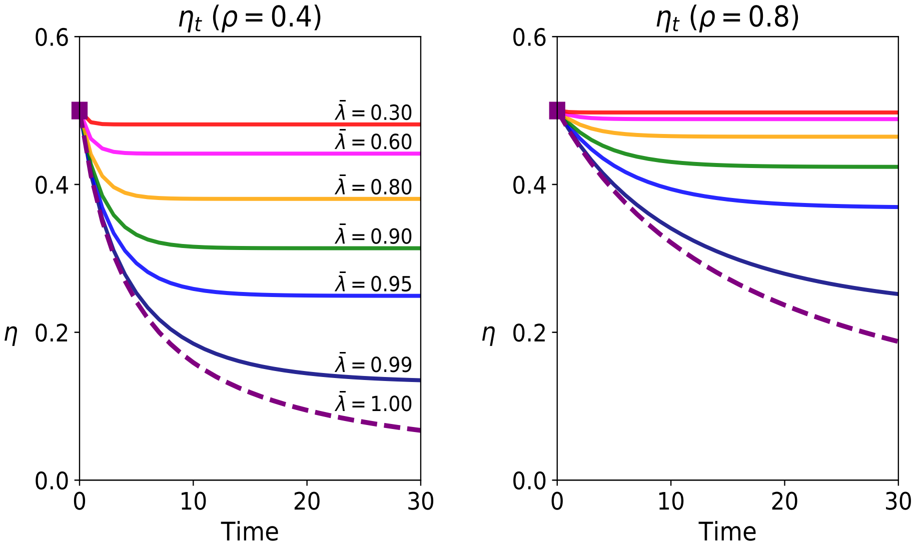


Figure 9: The evolution of scaled uncertainty about μ as the number t of previous (imperfectly remembered) observations grows. Each panel corresponds to a particular value of ρ (maintaining the assumption that $K = 1$, as in Figure 1). Each panel shows the evolution for several different possible values of $\bar{\lambda}$ (color code is the same in both panels).

$\eta_\infty = 0$ that is eventually reached. Instead, when memory is imperfect, the long-run value η_∞ is also higher when the state is more serially correlated; effectively, the limited number of recent observations of the state that can be retained in memory reveal less about the value of μ when the state is more serially correlated.

F.2 Solving for the value function $\tilde{V}(\eta)$ and policy function $\lambda^*(\eta)$ in the case of a linear information cost

In the case of a linear information cost (or any other cost function with a positive marginal cost of increasing I_t), it is necessary to solve the Bellman equation for the value function $\tilde{V}(\eta)$, in order to determine the optimal dynamics of $\{\lambda_t\}$. Here we explain the methods used to solve this problem in the case of a linear information cost (the results reported in section 3.2).

Once we have solved for the function $\phi(\eta_t; \lambda_t)$, as in the previous subsection, the Bellman equation for the case of a linear information cost can be written

$$\tilde{V}(\eta_t) = \min_{\lambda_t \in [0,1]} \left[\eta_t - \frac{\tilde{\theta}}{2} \log(1 - \lambda_t) + \beta \tilde{V}(\phi(\eta_t; \lambda_t)) \right]. \tag{F.3}$$

We use the value function iteration algorithm to find the value function that is a fixed point of this mapping.

When iterating the mapping to update the value function, we use a grid search method to find the optimal policy function, because the right-hand side of the Bellman equation is in general a non-convex function of the policy variable λ_t (as we illustrate in Figure 12 below).

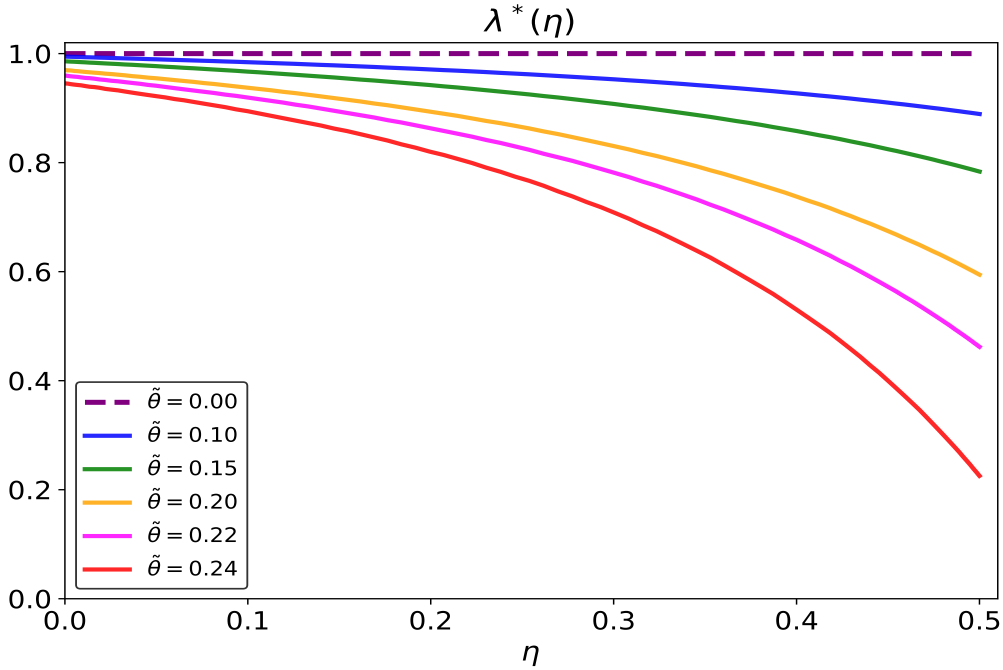


Figure 10: The optimal policy function $\lambda^*(\eta)$, in the case of progressively larger values for the information cost parameter $\tilde{\theta}$, under the assumption that $K = 1, \rho = 0$.

We approximate the value function with Chebyshev polynomials. Once the value function has converged, we can use our solution for $\tilde{V}(\eta)$ to solve numerically for the policy function $\lambda^*(\eta)$, the solution to the minimization problem on the right-hand side of (F.3).

This function is graphed for several values of $\tilde{\theta}$ in Figure 10, where we maintain the parameter values $K = 1, \rho = 0$ as in Figure 1. When $\tilde{\theta} = 0$ (no cost of memory precision), it is optimal to choose $\lambda_t = 1$ (perfect memory) in all cases. But for any value of η, the optimal $\lambda^*(\eta) < 1$ when $\tilde{\theta} > 0$ (since in this case, perfect memory becomes infinitely costly); furthermore it is lower (memory is more imperfect) the higher is $\tilde{\theta}$. We also see that for any cost parameter $\tilde{\theta} > 0$, the optimal $\lambda^*(\eta)$ is a decreasing function of η. This indicates that the less accurate the information contained in the cognitive state s_t (as indicated by the higher value of η_t), the less information about the cognitive state that it will be optimal to store in memory, when the memory cost can be reduced by storing a less informative record.

The policy function $\lambda_t = \lambda^*(\eta_t)$ together with the law of motion

$$\eta_{t+1} \;=\; \phi(\eta_t; \lambda_t) \tag{F.4}$$

derived in section F.1 can then be solved for the dynamics of the scaled uncertainty $\{\eta_t\}$ for all $t \geq 0$, starting from the initial condition $\eta_0 = K/(K+1)$. The dynamics implied by these equations can be graphed in a phase diagram, as illustrated in Figure 11. In the phase diagrams shown in each of the two panels, the value of η_t is indicated on the horizontal axis and the value of λ_t on the vertical axis. Equation (F.4), which holds regardless of the nature of the information cost function and the degree to which the future is discounted, determines a locus $\eta_\infty(\lambda)$, indicating for each value of λ the unique value of η that will be a fixed point of these dynamics if λ_t is held at the value λ. We can further show that whenever

Figure 4: The dynamics of scaled uncertainty and memory precision

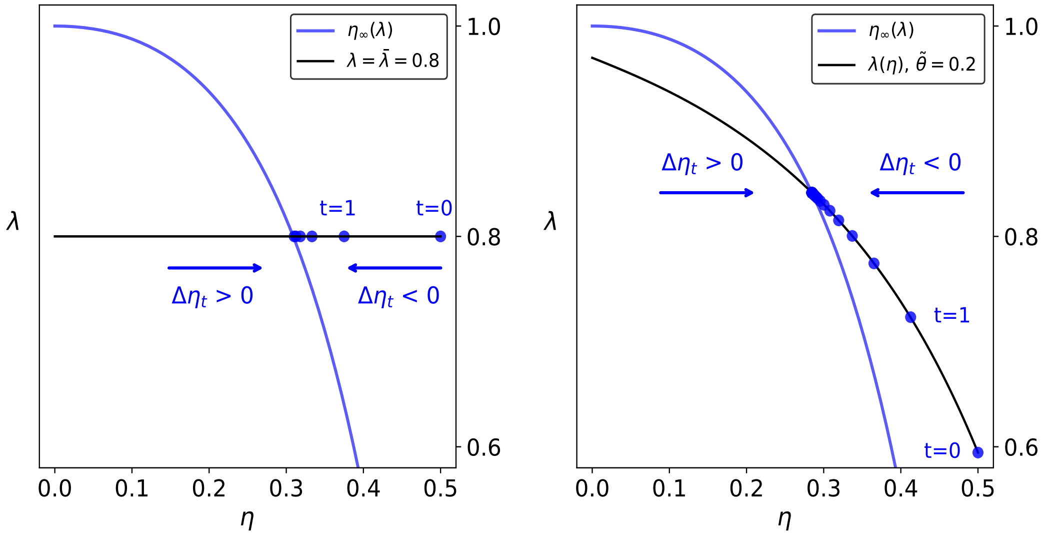


Figure 11: The dynamics of scaled uncertainty η_t and memory precision λ_t graphed in the phase plane. The left panel gives an alternative graphical presentation of the dynamics plotted in Figure 1 for the case of a fixed upper bound $\bar{\lambda}$ on memory precision. The right panel shows the corresponding dynamics in the case of a linear cost of precision parameterized by $\tilde{\theta}$.

$\eta_t < \eta_\infty(\lambda_t)$, the law of motion (F.4) implies that $\eta_{t+1} > \eta_t$, so that uncertainty will increase, while if $\eta_t > \eta_\infty(\lambda_t)$, it implies instead that $\eta_{t+1} < \eta_t$, so that uncertainty will decrease.

The choice of λ_t (and hence the degree to which uncertainty will increase or decrease) is given by the policy function, that depends on the specification of information costs. When there is a fixed upper bound on information (the case discussed in the previous subsection), the policy function is just a horizontal line at the vertical height $\bar{\lambda}$, as shown in the left panel of the figure.[74] In this case, the values of (η_t, λ_t) in successive periods start at the point $(\eta_0, \bar{\lambda})$, labeled "$t = 0$" in the figure, and then move left along the graph of the policy function (since $\eta_0 > \eta_\infty(\bar{\lambda})$ as shown). They continue to move left along the policy function, with η_t converging asymptotically to $\eta_\infty(\bar{\lambda})$ from above; the stationary long-run values $(\eta_\infty, \lambda_\infty)$ correspond to the point at which the policy function $\lambda = \bar{\lambda}$ intersects the locus of fixed points $\eta_\infty(\lambda)$.

The right-hand panel of the figure shows the corresponding phase-plane dynamics in the less trivial case of a linear cost function for information. In this case, the policy function is instead a downward-sloping curve, as shown in Figure 10.[75] Again the values of (η_t, λ_t) in successive periods must always lie on the graph of the policy function; the direction of

[74] The figure plots the location of this line for the case $\bar{\lambda} = 0.8$. The figure is drawn for parameter values $K = 1, \rho = 0$. Thus the dynamics of uncertainty shown in the figure correspond to the curve labeled $\bar{\lambda} = 0.8$ in Figure 1.

[75] In the figure, the policy function and the implied dynamics are shown for the case in which $\tilde{\theta} = 0.2$, corresponding to one of the intermediate curves shown in Figure 10. Again the figure is for the case $K = 1, \rho = 0$, so that the location of the locus of fixed points $\eta_\infty(\lambda)$ and the law of motion (F.4) remain the same as in the left panel.

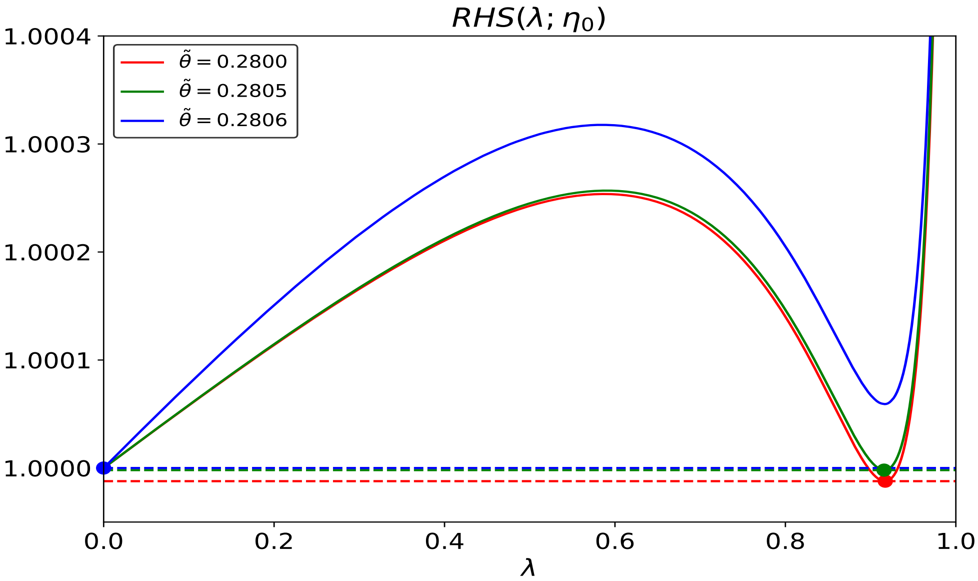


Figure 12: The objective function $RHS(\lambda_t, \eta_t)$ that is minimized in the Bellman equation, plotted as a function of λ_t for the initial level of uncertainty $\eta_t = \eta_0$. The function is normalized so that the value is 1.0 when $\lambda_t = 0$, and plotted for three nearby values of $\tilde{\theta}$, in the case that $K = 10$. The minimizing value of λ_t jumps discontinuously as $\tilde{\theta}$ passes a value between 0.2800 and 0.2805.

motion up or down this curve depends on whether the current position lies to the left or right of the locus of fixed points $\eta_\infty(\lambda)$. The initial point (labeled "$t = 0$") is determined as the point on the policy curve with horizontal coordinate given by the initial condition η_0. Since this point lies to the right of the locus of fixed points, the points for successive periods move up and to the left on the policy curve, meaning that λ_t rises as η_t falls.

The scaled uncertainty continues to fall, and the precision of memory continues to rise, until the values (η_t, λ_t) converge to stationary long-run values $(\eta_\infty, \lambda_\infty)$, again corresponding to the point at which the policy function $\lambda^*(\eta)$ intersects the locus of fixed points $\eta_\infty(\lambda)$. Note that convergence is slower in the right panel of the figure than in the left, because in the early periods, when uncertainty is high, a less precise memory is chosen in the linear-cost case, resulting in slower learning from experience.

Different values of $\tilde{\theta}$ correspond to different locations for the policy function $\lambda^*(\eta)$, as shown in Figure 10, and hence to different dynamics in the phase plane, converging to different long-run levels of scaled uncertainty. The dynamics of scaled uncertainty as a function of the number of observations t are shown for progressively larger values of $\tilde{\theta}$ in Figure 3 in the main text, using the same format as in Figure 1.

F.3 The possibility of discontinuous solutions

Figure 12 illustrates our comment about the possible non-convexity of the optimization problem (F.3). Let $RHS(\lambda_t; \eta_t)$ be the function defined on the right-hand side of (F.3), i.e., the objective of the minimization problem. The figure plots the value of $RHS(\lambda; \eta_0)$, normalized by dividing by the positive constant $RHS(0; \eta_0)$ (so that a value of 1.0 on the vertical axis means that $RHS(\lambda;\eta_0)$ is of exactly the same size as $RHS(0;\eta_0)$). This function is shown for each of three slightly different values of $\tilde{\theta}$, assuming in each case that $K = 10$, as in the right panel of Figure 5 in the text. In the case of each of these curves, a large dot (the same color as the curve) indicates the global minimum of the function. A horizontal dashed line (also the same color as the corresponding curve) indicates the minimum of $RHS(\lambda;\eta_0)$ — and thus the value of $\tilde{V}(\eta_0)$ — again normalized by dividing by $RHS(\eta_0)$.

The figure shows that for values of $\tilde{\theta}$ in this range, $RHS(\lambda)$ is not a convex function of λ. It is increasing for small enough values of λ, making the choice $\lambda_t = 0$ a local minimum in this case. (This is true for all values of $\tilde{\theta}$ greater than a critical value around 0.15, which explains the existence of the horizontal segment of the connected black curve in the right panel of Figure 5.) However, the function reaches a local maximum, and then decreases for larger values of λ, as the degree to which a larger value of λ_t reduces $\phi(\eta_0;\lambda_t)$ outweighs the increase in the information cost. (A large enough value of K is required for this to occur. A larger value of K increases the sensitivity of the value of $\phi(\eta_0;\lambda)$ to the value of λ; see equation (F.5) below.) For even larger values of λ (values approaching 1), further increases in λ increase the information cost term so sharply that $RHS(\lambda;\eta_0)$ is again decreasing in λ. This means that there is a second local minimum of the objective function, at an interior value of λ. Which of the two local minima represents the global minimum of the function depends on parameter values.

In the case illustrated in the figure, the interior local minimum achieves a lower value of the objective than the choice $\lambda_t = 0$, for all values of $\tilde{\theta}$ less than a critical value that is slightly larger than 0.2805. (As shown in the figure, when $\tilde{\theta} = 0.2805$, the interior minimum achieves a value of the objective that is quite close to the value $RHS(0;\eta_0)$. However, the value achieved remains slightly smaller: there is a (barely visible) green dashed line, just below the blue dashed line at the normalized value 1.0.) But the normalized value of the objective at the interior minimum increases as $\tilde{\theta}$ is increased, and for a value of $\tilde{\theta}$ only slightly greater than 0.2805, the normalized value becomes greater than 1.0 (which is to say, the interior local minimum is no longer the global minimum of the objective). When this critical value of $\tilde{\theta}$ is passed, the optimal value $\lambda^*(\eta_0)$ jumps discontinuously from the interior local minimum (which is a continuously decreasing function of $\tilde{\theta}$) to the value zero. When this happens, the optimal long-run level for the normalized uncertainty measure η_∞ increases discontinuously, from a value on the lower branch of the correspondence shown in the right panel of Figure 5 to the value $\eta_0 = K/K + 1$. For all values of $\tilde{\theta}$ higher than this, it is optimal to choose a completely uninformative memory for all t, so that $\eta_t = \eta_0$ for all t, and hence $\eta_t \to \eta_\infty = \eta_0$.

For larger values of $\tilde{\theta}$ than those considered in Figure 10, the optimal policy function $\lambda^*(\eta)$ is equal to zero for all large enough (though still finite) values of η, as illustrated in Figure 13. Once $\tilde{\theta}$ is large enough for $\lambda^*(\eta_0)$ to equal zero, the optimal dynamics imply $\eta_t = \eta_0$ for all t, and hence $\eta_\infty = \eta_0 = K/K + 1$, as shown in Figure 5.

F.4 The case $\rho = 0$

Additional analytical results are possible in the case that $\rho = 0$ (the external state is an i.i.d. random variable). In this case, the law of motion for the scaled uncertainty measure

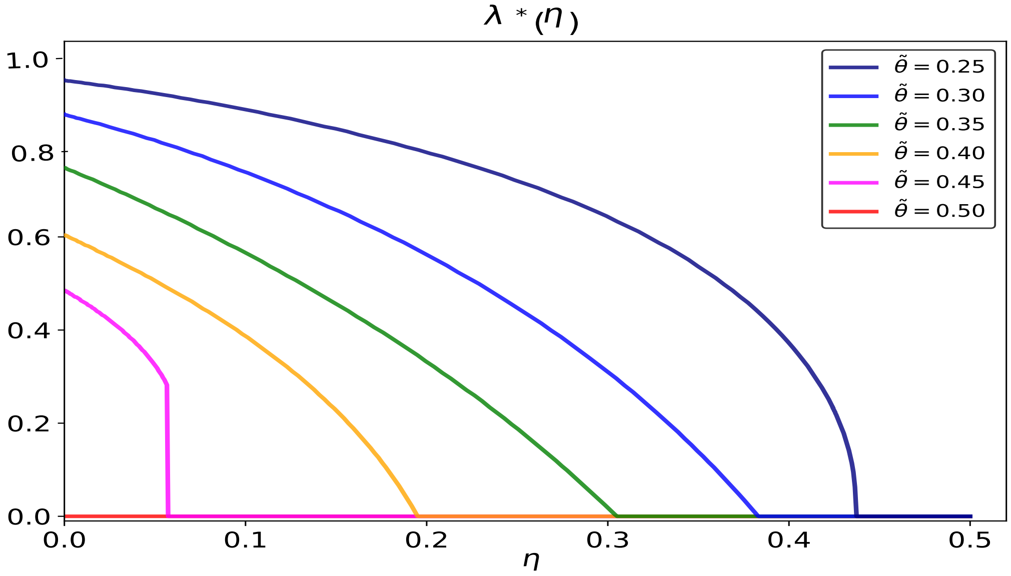


Figure 13: The optimal policy function $\lambda^*(\eta)$, in the case of progressively larger values for the information cost parameter $\tilde{\theta}$, under the assumption that $K = 1, \rho = 0$. Here we consider values of $\tilde{\theta}$ larger than those shown in Figure 10.

(derived in section F.1) simplifies to

$$\eta_{t+1} = 1 - \frac{1}{K + 1 - \lambda_t(K - \eta_t)} \;\equiv\; \phi(\eta_t; \lambda_t). \tag{F.5}$$

In the case of an exogenous upper bound on mutual information, the nonlinear difference equation obtained by setting $\lambda_t = \bar{\lambda}$ in (F.5) is of an especially simple sort. The function on the right-hand side of this equation is a hyperbola, increasing and concave for all $\eta_t > 0$. We easily see that the right-hand side has a positive value when $\eta_t = 0$, and a value less than $K/(K+1)$ when $\eta_t = K/(K+1)$.

Thus for any $0 < \bar{\lambda} < 1$, the function $\phi(\eta_t; \bar{\lambda})$ is an increasing, concave function that is above the diagonal at $\eta_t = 0$ and below the diagonal at $\eta_t = K/(K+1)$. It follows that the function must intersect the diagonal at exactly one point, $\eta_t = \eta_\infty$. We can furthermore give an explicit algebraic solution for this fixed point as the solution to a quadratic equation. Note in particular that it is necessarily strictly positive and strictly less than $K/(K+1)$, and that it is a decreasing function of $\bar{\lambda}$, approaching $K/(K+1)$ as $\bar{\lambda} \to 0$, and approaching 0 as $\bar{\lambda} \to 1$.

On the interval $\eta_\infty < \eta_t \leq K/(K+1)$, the law of motion (F.5) implies that $\eta_\infty < \eta_{t+1} < \eta_t$. Hence when we start from the initial condition $\eta_0 = K/(K+1)$, the implied dynamics must satisfy

$$\eta_0 \;>\; \eta_1 \;>\; \eta_2 \;>\; \eta_3 \;\ldots,$$

a monotonically decreasing sequence. Because the sequence is bounded below by η_∞, it must converge, and it is easily seen that it can only converge to the fixed point η_∞ that we have already calculated. Hence for each possible $\bar{\lambda}$, we obtain a monotonically decreasing, convergent sequence of the kind shown in Figure 1. We can also easily show that the curve must be lower for each value of t, the larger is $\bar{\lambda}$.

We can also obtain additional analytical results in the case of a linear information cost. The value function satisfies a Bellman equation of the form

$$\tilde{V}(\eta_t) = \min_{\lambda_t} \left[\beta^2 \eta_t - \frac{\tilde{\theta}}{2} \log\left(1 - \lambda\right) + \beta \tilde{V}\left(\phi(\eta_t; \lambda_t)\right)\right].$$

The first order condition with respect to λ_t is

$$\frac{\tilde{\theta}}{2} \frac{1}{1 - \lambda_t} + \beta \tilde{V}'(\eta_{t+1}) \frac{\partial \phi(\eta_t; \lambda_t)}{\partial \lambda_t} = 0. \tag{F.6}$$

And the envelope condition is

$$\tilde{V}'(\eta_t) = \beta^2 + \beta \tilde{V}'(\eta_{t+1}) \frac{\partial \phi(\eta_t; \lambda_t)}{\partial \eta_t}.$$

We can use these two conditions to derive an Euler equation for the dynamics of the scaled uncertainty measure.

Substituting the solution (F.5) for $\phi(\eta_t; \lambda_t)$ and taking the derivative with respect to λ_t, we can rewrite (F.6) as

$$\begin{aligned}
\tilde{V}'(\eta_{t+1}) &= -\frac{\tilde{\theta}}{2\beta} \frac{1}{1 - \lambda_t} \left(\frac{\partial \phi(\eta_t; \lambda_t)}{\partial \lambda_t} \right)^{-1} \\
&= -\frac{\tilde{\theta}}{2\beta} \frac{1}{1 - \lambda_t} \left(-\frac{(K - \eta_t)}{\left(K + 1 - \lambda_t (K - \eta_t)\right)^2} \right)^{-1} \\
&= \frac{\tilde{\theta}}{2\beta} \frac{\left(K + 1 - \lambda_t (K - \eta_t)\right)^2}{(1 - \lambda_t)(K - \eta_t)} \\
&= \frac{\tilde{\theta}}{2\beta} \frac{1}{\left(1 - \eta_{t+1}\right)\left(1 - (1 - \eta_{t+1})(1 + \eta_t)\right)},
\end{aligned}$$

where the last equality is derived by again substituting the law of motion (F.5). It follows that if $\eta_t \to \eta_\infty$ in the long run, the stationary solution η_∞ must satisfy

$$\tilde{V}'(\eta_\infty) = \frac{\tilde{\theta}}{2\beta} \frac{1}{(1 - \eta_\infty)\eta_\infty^2}. \tag{F.7}$$

Next we rewrite (F.4), again taking the derivative of expression (F.5) for $\tilde{}(\eta_t; \lambda_t)$:

$$\begin{aligned}
\tilde{V}'(\eta_t) &= \beta^2 + \beta \tilde{V}'(\eta_{t+1}) \frac{\partial \phi(\eta_t; \lambda_t)}{\partial \eta_t} \\
&= \beta^2 + \beta \tilde{V}'(\eta_{t+1}) \frac{\lambda_t}{\left(K + 1 - \lambda (K - \eta_t)\right)^2} \\
&= \beta^2 + \beta \tilde{V}'(\eta_{t+1}) \frac{\lambda_t}{\left(1 - \eta_{t+1}\right)^{-2}} \\
&= \beta^2 + \beta \tilde{V}'(\eta_{t+1}) (1 - \eta_{t+1})^2 \frac{(K + 1)(1 - \eta_{t+1}) - 1}{(K - \eta_t)(1 - \eta_{t+1})}.
\end{aligned}$$

It follows that the stationary solution η_∞ must satisfy

$$\tilde{V}'(\eta_\infty) = \beta^2 + \beta \tilde{V}'(\eta_\infty) \frac{(1-\eta_\infty)\left[(K+1)(1-\eta_\infty)-1\right]}{K-\eta_\infty}. \tag{F.8}$$

Moreover, in a stationary solution, the value $\tilde{V}'(\eta_\infty)$ given by (F.7) must also be the value of $\tilde{V}'(\eta_\infty)$ in (F.8). Using (F.7) to substitute for $\tilde{V}'(\eta_\infty)$ in (F.8), we obtain a condition that must be satisfied by η_∞ in any stationary solution with an interior optimum (i.e., a stationary solution in which $0 < \eta_\infty < K/(K+1)$):

$$\tilde{\theta} = 2\beta^3(1-\eta_\infty)\eta_\infty^2 \left[1 - \beta \frac{(K+1)(1-\eta_\infty)^2 - (1-\eta_\infty)}{K-\eta_\infty}\right]^{-1}. \tag{F.9}$$

This is the relationship between $\tilde{\theta}$ and η_∞ that is graphed as a connected black curve in Figure 5. Note that for any value $0 < \eta_\infty < K/(K+1)$, there is a unique $\tilde{\theta} > 0$ consistent with this relationship; but (as shown in the right panel of Figure 5) there may be multiple solutions for η_∞ consistent with a given value of $\tilde{\theta}$.

G Predicted Values for the Quantitative Measures of Forecast Bias

Here we provide further explanation of the numerical results reported in section 4 of the main text.

G.1 Long-run stationary fluctuations

From the definition of the univariate memory state $\tilde{m}_{t+1} = \lambda_t v_t' \bar{s}_t + \omega_{t+1}$, we can derive a law of motion for the univariate memory state $\tilde{m}_t$. Using the subscript ∞ for the long-run stationary coefficients, we get

$$\begin{aligned}
\tilde{m}_{t+1} &= \lambda_\infty v_\infty' \bar{s}_t + \tilde{\omega}_{t+1} \\
&= \lambda_\infty v_\infty' \begin{pmatrix} \hat{\mu}_t \\ y_t \end{pmatrix} + \tilde{\omega}_{t+1} \\
&= \lambda_\infty \left[e_1' v_\infty \left\{ (e_1' - \gamma_1 c') m_t + \gamma_1 y_t \right\} + (e_2' v_\infty) y_t \right] + \tilde{\omega}_{t+1} \\
&= \lambda_\infty \left[e_1' v_\infty \left\{ (e_1' - \gamma_1 c') X_\infty v_\infty \tilde{m}_t + \gamma_1 y_t \right\} + (e_2' v_\infty) y_t \right] + \tilde{\omega}_{t+1} \\
&= \rho_m \tilde{m}_t + \rho_{my} y_t + \tilde{\omega}_{t+1}
\end{aligned}$$

where $\rho_m \equiv \lambda_\infty (e_1' v_\infty)\,(e_1' - \gamma_1 c')\, X_\infty v_\infty$ and $\rho_{my} \equiv \lambda_\infty\,(\gamma_1 + e_2' v_\infty)$.

We can evaluate the numerical values of the coefficients defining the long-run dynamics as follows. Equations (F.1)–(F.2) imply that the long-run coefficients $\lambda_\infty, \eta_\infty, \gamma_{1,\infty}$ must satisfy the pair of nonlinear equations

$$\eta_\infty \;=\; \frac{(1-\lambda_\infty)(1-\gamma_{1,\infty})^2 K \;+\; (1-\rho^2\lambda_\infty)\gamma_{1,\infty}^2}{1 \;-\; \lambda_\infty(1-(1-\rho)\gamma_{1,\infty})^2},$$

$$\gamma_{1,\infty} \;=\; \frac{(1-\lambda_\infty)K + (1-\rho)\lambda_\infty\eta_\infty}{(1-\lambda_\infty)(K+\rho^2) + (1-\rho^2) + (1-\rho)^2\lambda_\infty\eta_\infty}.$$

In the case of an exogenous bound on mutual information, we can set $\lambda_\infty = \bar{\lambda}$, in which case these provide two equations to solve for the values of η_∞ and $\gamma_{1,\infty}$. (Note that the relevant solution is the one that satisfies the bounds $0 < \eta_\infty < K/(K+1)$, and that it necessarily also satisfies $0 < \gamma_{1,\infty} < 1/(1-\rho)$.) This allows us to compute the long-run stationary values of the coefficients η and γ_1 plotted for alternative values of $\bar{\lambda}$ in Figure 2.

We have also shown in section E.3 that the optimal weight vector v_t is just a normalized version of the vector $\delta_{t+1} \equiv e_1 - \gamma_{1,t+1}c$. Hence in the long run, this vector must become

$$v_\infty \;=\; \frac{e_1 \,-\, \gamma_{1,\infty}c}{(e_1' - \gamma_{1,\infty}c')X_\infty(e_1 - \gamma_{1,\infty}c)}.$$

In particular, the ratio $v_{2,\infty}/v_{1,\infty}$ (the quantity plotted as "v_∞" in Figure 2) is given by

$$\frac{v_{2,\infty}}{v_{1,\infty}} \;=\; -\frac{\rho\gamma_{1,\infty}}{1 \,-\, (1-\rho)\gamma_{1,\infty}} \;<\; 0.$$

Finally, we observe that the intrinsic persistence coefficient ρ_m defined above must satisfy

$$\begin{aligned} \rho_m \;&\equiv\; \lambda_\infty v_{1,\infty} \cdot (e_1' - \gamma_{1,\infty}c')X_\infty v_\infty \\ &=\; \lambda_\infty v_{1,\infty} \\ &=\; \lambda_\infty(1-(1-\rho)\gamma_{1,\infty}). \end{aligned}$$

This allows us to calculate the other coefficient that is plotted in Figure 2. Note that because the Kalman gain necessarily satisfies the bounds $0 < \gamma_1 < 1/(1-\rho)$, this solution for the intrinsic persistence coefficient implies that

$$0 \;<\; \rho_m \;<\; 1. \tag{G.1}$$

In the long run, we can describe the evolution of the DM's cognitive state using the following system of equations:

$$\begin{aligned} \tilde{m}_{t+1} &= \rho_m\tilde{m}_t + \rho_{my}y_t + \tilde{\omega}_{t+1} \\ y_{t+1} &= (1-\rho)\mu + \rho y_t + \epsilon_{y,t+1} \end{aligned}$$

Therefore, we can write it as a VAR(1) system with constant coefficients and Gaussian innovation terms:

$$\begin{pmatrix}\tilde{m}_{t+1}\\ y_{t+1}\end{pmatrix} = \begin{pmatrix}0\\ 1-\rho\end{pmatrix}\mu + \begin{pmatrix}\rho_m & \rho_{my}\\ 0 & \rho\end{pmatrix}\begin{pmatrix}\tilde{m}_t\\ y_t\end{pmatrix} + \begin{pmatrix}\tilde{\omega}_{t+1}\\ \epsilon_{y,t+1}\end{pmatrix}$$

Because the two eigenvalues of this vector law of motion are ρ and ρ_m, (G.1) implies that this describes a stationary stochastic process. Hence we can compute stationary long-run values for the second moments of the variables, and use these to define the impulse response functions and predicted regression coefficients reported in the text.

For example, in the case of a fixed per-period bound on mutual information, we can compute the impulse responses for the DM's estimate of μ and her one-quarter-ahead forecast of the external state, as explained in section 3.3. Here we present additional figures, showing what the impulse responses shown in Figure 6 in the text would be like in the case of alternative values of ρ. In Figures 14 and 15 shown here, each panel corresponds to a different value of ρ, and shows the responses for several different possible values of $\bar{\lambda}$. (As with Figure 6 in the main text, we here assume that $K = 1$.)

Figure 14: Impulse responses of the DM's estimate of μ for alternative degrees of persistence ρ of the external state process.

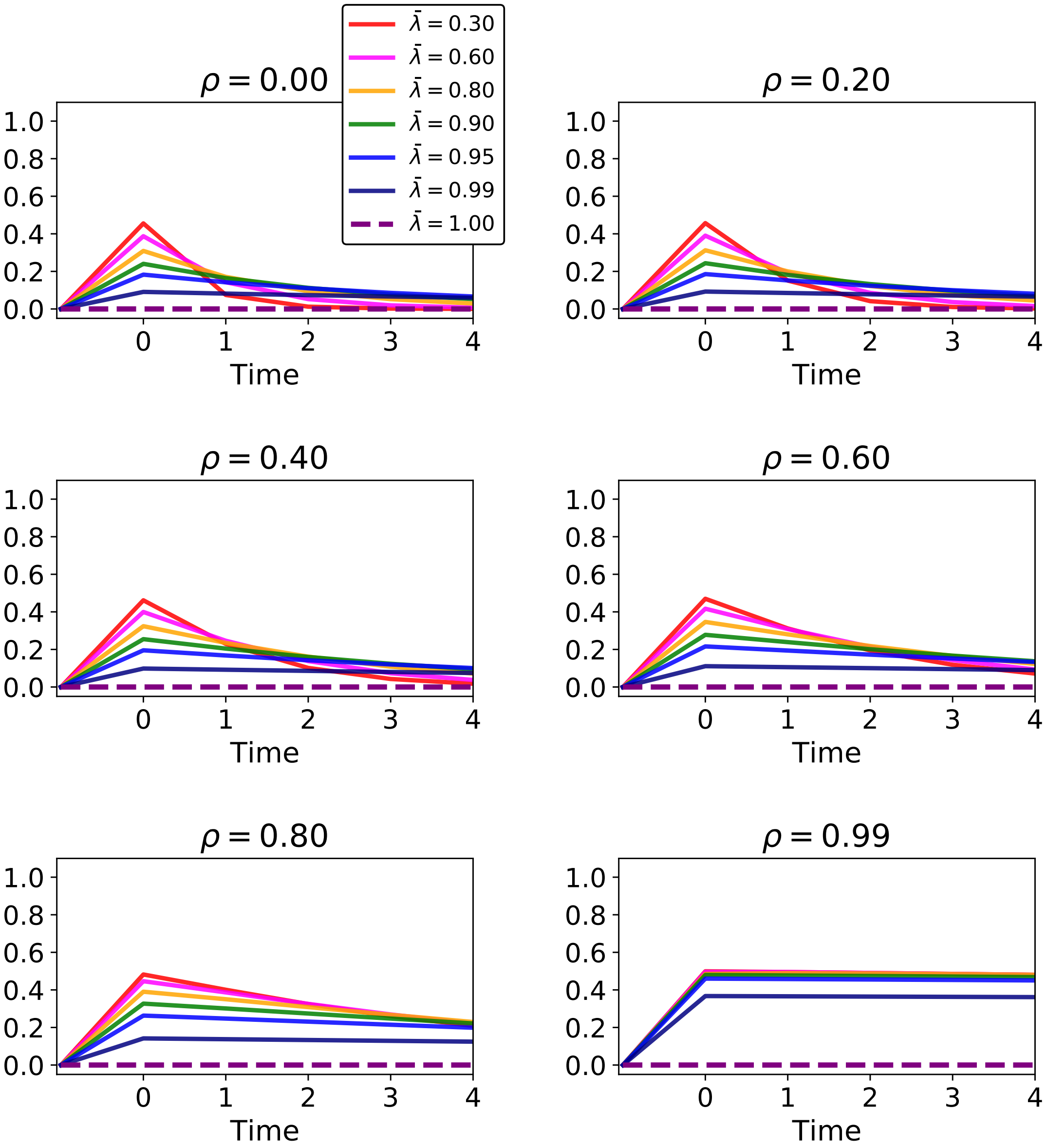
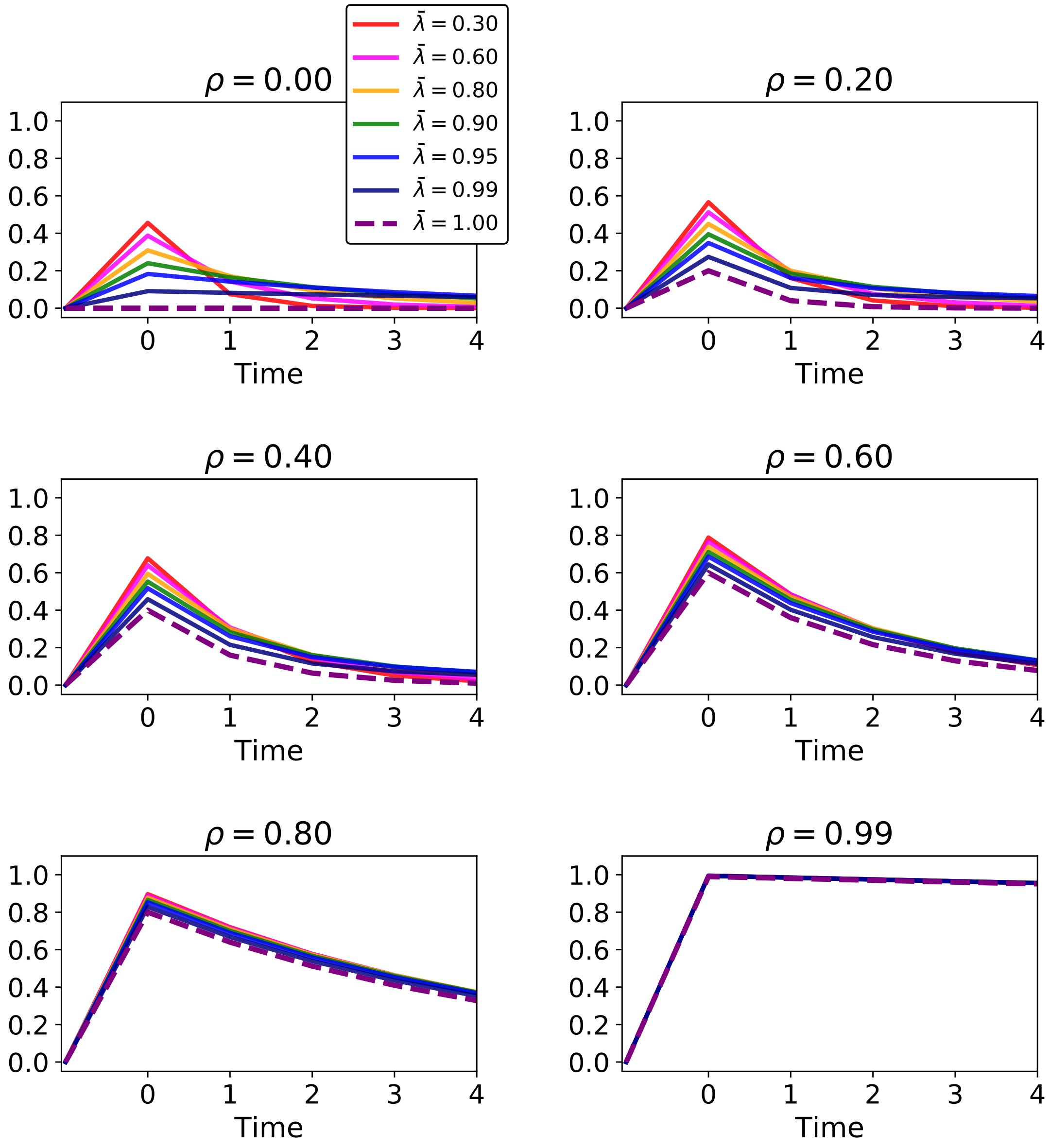


Figure 15: Impulse responses of the DM's one-quarter-ahead forecast of the external state for alternative degrees of persistence ρ of the external state process.



G.2 Predicted value of the regression coefficient ρ_h^{subj}

Given a long enough series of observations from an environment with a fixed μ, our model yields stationary values for the Kalman gain γ_1 and for the amplitude of fluctuations in the memory state $var[\bar{m}_t]$. We can then compute the values of the following long-run conditional second moments:

$$\begin{aligned} var[\bar{m}_t|\mu] &= var[\bar{m}_t] - cov[\bar{m}_t, \mu] var[\mu]^{-1} cov[\mu, \bar{m}_t] \\ &= var[\bar{m}_t] - cov[\bar{m}_t, x_t] e_1 var[\mu]^{-1} e_1' cov[x_t, \bar{m}_t] \\ &= var[\bar{m}_t] - \frac{1}{var[\mu]} var[\bar{m}_t] e_1 e_1' var[\bar{m}_t] \end{aligned}$$

$$\begin{aligned} cov[\hat{\mu}_t, y_t|\mu] &= cov[(e_1' - \gamma_1 c')\bar{m}_t + \gamma_1 y_t, y_t|\mu] \\ &= (e_1' - \gamma_1 c') cov[\bar{m}_t, y_t|\mu] + \gamma_1 var[y_t|\mu] \\ &= (e_1' - \gamma_1 c') var[\bar{m}_t|\mu] c + \gamma_1 var[y_t|\mu] \end{aligned}$$

$$\begin{aligned} var[\hat{\mu}_t|\mu] &= var[(e_1' - \gamma_1 c')\bar{m}_t + \gamma_1 y_t|\mu] \\ &= (e_1' - \gamma_1 c') var[\bar{m}_t|\mu](e_1 - \gamma_1 c) + \gamma_1^2 var[y_t|\mu] + 2\gamma_1 (e_1' - \gamma_1 c') cov[\bar{m}_t, y_t|\mu] \\ &= (e_1' - \gamma_1 c') var[\bar{m}_t|\mu](e_1 - \gamma_1 c) + \gamma_1^2 var[y_t|\mu] + 2\gamma_1 (e_1' - \gamma_1 c') var[\bar{m}_t|\mu] c \end{aligned}$$

In order to write the dynamics of the model in terms of scale-invariant quantities, we divide each second moment by $var[y_t|\mu] = \sigma_y^2$. Thus we can write

$$\begin{aligned} \frac{var[\bar{m}_t|\mu]}{var[y_t|\mu]} &= \tilde{\Sigma}_{\bar{m}} - \frac{1}{K} \tilde{\Sigma}_{\bar{m}} e_1 e_1' \tilde{\Sigma}_{\bar{m}} \\ \frac{cov[\hat{\mu}_t, y_t|\mu]}{var[y_t|\mu]} &= (e_1' - \gamma_1 c') \frac{var[\bar{m}_t|\mu]}{var[y_t|\mu]} c + \gamma_1 \\ \frac{var[\hat{\mu}_t|\mu]}{var[y_t|\mu]} &= (e_1' - \gamma_1 c') \frac{var[\bar{m}_t|\mu]}{var[y_t|\mu]} (e_1 - \gamma_1 c) + \gamma_1^2 + 2\gamma_1 (e_1' - \gamma_1 c') \frac{var[\bar{m}_t|\mu]}{var[y_t|\mu]} c, \end{aligned}$$

using the notation $\tilde{\Sigma}_{\bar{m}} \equiv \mathrm{var}[\bar{m}_t]/\sigma_y^2$.

We now wish to calculate the predicted asymptotic value of the regression coefficient

$$\rho_h^{subj} \equiv \frac{cov[\hat{y}_{t+h|t}, y_t|\mu]}{var[y_t|\mu]}$$

where $\hat{y}_{t+h|t} \equiv E[y_{t+h}|\bar{m}_t, y_t]$. From

$$\begin{aligned} cov[\hat{y}_{t+h|t}, y_t|\mu] &= cov[(1-\rho^h)\hat{\mu}_t + \rho^h y_t, y_t|\mu] \\ &= (1-\rho^h) cov[\hat{\mu}_t, y_t|\mu] + \rho^h var[y_t|\mu], \end{aligned}$$

where $\hat{\mu}_t \equiv E[\mu|\bar{m}_t, y_t]$, we can then compute

$$\begin{aligned} \rho_h^{subj} &= (1-\rho^h) \frac{cov[\hat{\mu}_t, y_t|\mu]}{var[y_t|\mu]} + \rho^h \\ &= (1-\rho^h) \left[(e_1' - \gamma_1 c') \left(\tilde{\Sigma}_{\bar{m}} - \frac{1}{K} \tilde{\Sigma}_{\bar{m}} e_1 e_1' \tilde{\Sigma}_{\bar{m}} \right) c + \gamma_1 \right] + \rho^h. \end{aligned}$$

These are the coefficients whose values are plotted against the value of $\rho_h = \rho^h$ in Figure 7.