A Model of Costly Recall

Hassan Afrouzi* Columbia
Spencer Kwon† Harvard
Yueran Ma‡ Chicago Booth

April 10, 2020
Preliminary Draft

Abstract

Recent evidence suggests that beliefs overreact to news relative to the frictionless rational benchmark. Moreover, the degree of overreaction appears to be more pronounced when the underlying process has lower persistence or when the forecast horizon is longer. In this paper, we propose a theory of expectations formation with noisy recall from memory, which provides predictions in line with these facts. We estimate our model based on one-period-ahead forecasts, and find that the model outperforms commonly used models in matching key features of the data. The model also matches longer horizon forecasts as non-targeted moments.

JEL classification: D84; D91

Key Words: Memory, Belief Formation, Noisy Recall

*hassan.afrouzi@columbia.edu
†ykwon@hbs.edu
‡yueran.ma@chicagobooth.edu

1 Introduction

How expectations respond to news is central for understanding economic decision making. There is accumulating evidence that expectations display systematic deviations from frictionless rational benchmarks. However, empirical findings on the form of deviations have been mixed. Some studies point to overreaction: beliefs update too much with respect to news, or imply subjective persistence higher than the true persistence of the process (De Bondt and Thaler, 1990; Bordalo, Gennaioli and Shleifer, 2018a; Bordalo, Gennaioli, Ma and Shleifer, 2019c; Barrero, 2018). Meanwhile, other studies show underreaction and insufficient adjustment (Abarbanell and Bernard, 1992; Bouchaud, Krueger, Landier and Thesmar, 2019; Ma, Ropele, Sraer and Thesmar, 2020). Recent work across different settings suggests two potential ways to connect these disparate patterns. First, overreaction appears more pronounced when the underlying process has lower persistence (Bordalo, Gennaioli, Ma and Shleifer, 2019c; Landier, Ma and Thesmar, 2020). Second, overreaction also appears more persistent when the forecast horizon is longer (Giglio and Kelly, 2018; Bordalo, Gennaioli, La Porta and Shleifer, 2019b; Wang, 2019).

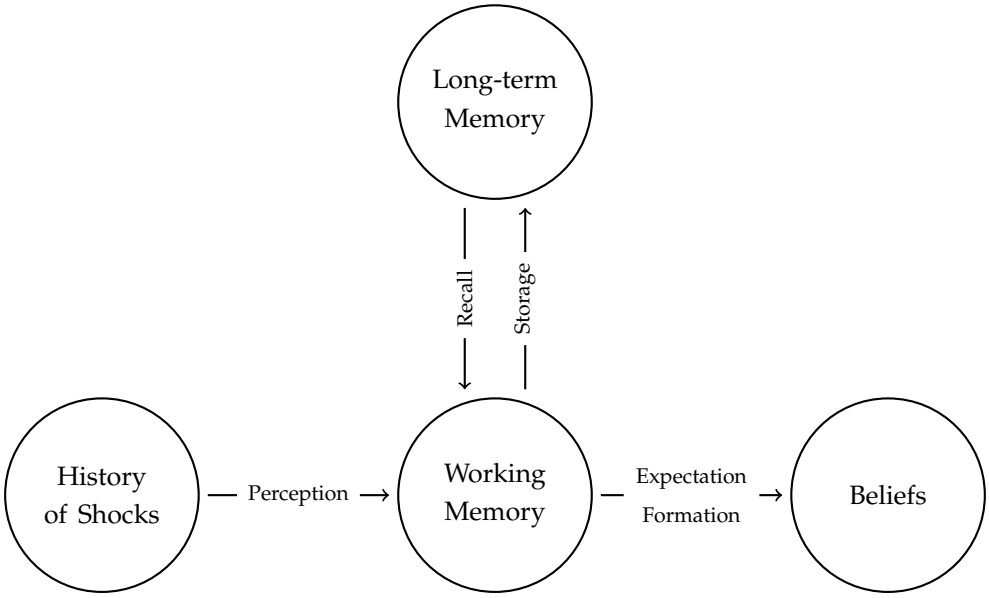
While there is a wide variety of models that seek to explain the systematic biases in expectations formation, these models have difficulty generating the two patterns above in terms of how biases vary depending on process persistence and forecast horizon, as discussed in Landier, Ma and Thesmar (2020). We propose a new model of expectations formation with noisy recall from memory, which fits well with these two general observations.

In our model, the agent initially observes a context, such as the most recent realization of a process, which automatically forms the initial prior. Then, the agent decides how much to recall relevant information from memory to form posterior beliefs, subject to a cost of retrieval. Consequently, in our framework, the agent's final expectations reflect a balance between current news and memories of past experiences.

Formally, if we define "working memory" to be the set of signals that shape the agent's beliefs, three separate processes interact for belief formation: perception, storage, and recall (Figure 1). The first process, perception, governs how news is encoded in the agent's working memory. The second process, storage, governs how past news is stored in memory for later use. The third process, recall, governs how stored information is retrieved to working memory in order to contribute to belief formation.

As a benchmark, full information rational expectations (FIRE) is a model that has no friction associated with any of these channels: shocks are perceived (observed) perfectly by the agent, stored forever in memory and retrieved perfectly on demand. In such a model, the reaction to news is perfectly balanced from the perspective of an omniscient agent who has observed all realized shocks. This creates a natural benchmark for the strength of re-

Figure 1: A Conceptual Map of Belief Formation



Notes: The figure shows a conceptual map of the three processes for belief formation: news is *perceived* to working memory, *stored* in the long-term memory and *recalled* to contribute to beliefs.

action to news: an agent that reacts less strongly to news than an agent with FIRE exhibits "underreaction," while an agent that reacts more strongly to news exhibits "overreaction."

A large part of the literature on information frictions aims to explain shortcomings of FIRE by imposing frictions or costs through the perception channel. For instance, noisy information models (e.g., Woodford (2003)) or rational inattention models (e.g., Sims (2003)) assume frictions on the perception channel while leaving the storage and recall processes untouched. However, these models are generally inconsistent with the evidence documenting overreaction of beliefs to news and overpersistence of subjective beliefs.

Our main contribution in this paper is to study the implications of a different type of friction which operates through the recall process, rather than through the perception process. In particular, we propose a theory of costly recall that generates endogenous overpersistence in expectations as well as overreaction to news. In our model, the agent perceives shocks perfectly and has to form beliefs about the future long-run mean of a stochastic process. If recall is costless, the agent can perfectly obtain it from past realizations. Recall of past observations, however, is costly, which leads the agent to balance the precision of recall with recall costs. Accordingly, beliefs become disproportionately affected by the current realization of the shock.

The key mechanism in the model is the trade-off between *perception* of current shocks (news) and *recall* of past experiences and memories (anything but news). An agent who

has perfect recall of the past, but noisy perception of the current shocks, will always rely more on past memories and *underreact* to news. This is at the core of the failure of many models to capture overreaction to news. As long as perception is more costly than recall, the agent will discount perceived news in favor of past memories for the objective of his belief formation. On the other hand, in our framework, we emphasize that recall is noisy and costly relative to perception. Consequently, the agent will rely more on the current shock, leading to overreaction to news, in a manner similar to da Silveira and Woodford (2019). In particular, the agent's assessment of the long-run mean is swayed by the recent observations. Furthermore, our model delivers the novel prediction that the degree of overreaction depends positively on horizon and negatively on persistence: intuitively, information about the long-run mean is more valuable for predicting more transient, or longer-term outcomes.

We test our model using the experimental evidence on expectations formation from Landier, Ma and Thesmar (2020). In the experiment, participants observe past realizations of an AR(1) process and are asked to predict its future values. The main finding of Landier, Ma and Thesmar (2020) is that the forecasts show a higher level of subjective persistence than the true persistence of the underlying AR(1) process, and the degree of "overpersistence" bias is larger when the true persistence is lower. Landier, Ma and Thesmar (2020) also find that most existing models of expectations formation do not easily match the behavior of subjective persistence in the data. Nonetheless, we find that our micro-founded model of recall matches the key patterns in the data very well.

Furthermore, our model predicts that the degree of overpersistence in expectations should increase with the forecast horizon. We estimate our model based on one period ahead forecasts, and then compute the model-implied forecasts for other horizons. In this case, we also find that the model matches the data on longer horizon forecasts as non-targeted moments.

Overall, the key for the model's success in generating more overreaction for less persistent processes and longer forecast horizons, which is often observed in the data, follows from the mechanisms explained above. With costly recall, the agent's estimate of the long-run mean is disproportionately influenced by the most recent observations, which results in overreaction. The long-run mean is more relevant when the process is less persistent, or when the forecast horizon is longer, leading to more pronounced overreaction.

Literature review. Our paper is related to the literature on noisy perception and rational inattention (Woodford, 2003; Sims, 2003). While this literature focuses on the frictions in the perception channel of belief formation, our model emphasizes the frictions in the recall channel. In this sense, our model is related to da Silveira and Woodford (2019), where agents optimize over their memory subject to a cost. While they focus mainly on the implications of noisy memory for overreaction, we use our model to highlight how overreaction and

overpersistence of beliefs vary with the setting such as the persistence of the process and the horizon of forecasts, which are key issues for unifying empirical findings.

Our work also relates to the growing literature on memory and belief formation. Several papers explore implications of the laws of judgment and memory found in psychology research. Bordalo, Gennaioli and Shleifer (2018a), Bordalo, Gennaioli and Shleifer (2020), and Bordalo, Coffman, Gennaioli, Schwerter and Shleifer (2019a) draw inspirations from representativeness (Kahneman and Tversky, 1972) and associative recall (Kahana, 2012). Wachter and Kahana (2019) present a retrieved-context theory for belief formation under uncertainty. Our model incorporates both instinctive associative recall shaped by laws of memory and conscious recall that comes at a cost, and we allow agents to balance these two forces depending on the setting.

2 Model

Payoffs. Time is discrete and is indexed by $t \in \{0, 1, 2, \dots\}$. There is an agent who, at any time t, chooses an action vector $\vec{a}_t \in \mathbb{R}^m$ and gains an instantaneous payoff of $v(\vec{a}_t; \vec{x}_t)$ where $\{\vec{x}_t \in \mathbb{R}^{n_x}\}_{t\geq 0}$ is an stochastic process. Furthermore, we assume that the agent does not necessarily observe $\vec{x}_t$ directly but observes another process $\{\vec{y}_t \in \mathbb{R}^{n_y}\}$ that is potentially informative of $\vec{x}_t$.

As an example, $\vec{x}_t$ could denote the permanent component of an agent's income and $\vec{y}_t$ the contemporaneous income the agent that contains his transitory income shocks.

Perception and Storage. Perception and storage are both costless for the agent, meaning that at any time t the agent observes the realized $\vec{y}_t$ and stores it in his long-term memory. Therefore, the agent's long-term memory bank is equal to the history of all observed shocks,

$$Y^t \equiv (\vec{y}_s)_{s\leq t}. \tag{2.1}$$

Working Memory. Agent's beliefs are formed and actions are chosen conditional on his working memory, S_t. Due to costless perception, S_t always contains the realized value of $\vec{y}_t$ at time t as the agent has to perceive it in that period. Accordingly we refer to $\vec{y}_t$ as the *context* at t.

S_t can also contain other realized values of $\vec{y}_t$ based on two separate systems:

System 1: Associative [Unconscious] Recall. We allow for $\vec{y}_t$ to pull memories from the memory bank for free according to a "natural" rule. This assumption allows us to model the physiological properties of recall such as associative recall or recency effect. We interpret

these processes to be governed by the workings of the human brain and to be independent of conscious efforts of the agent.

Formally, we let $R_a(\vec{y}_t)$ denote the associative recall set based on $\vec{y}_t$. Note $\vec{y}_t \subseteq R_a(\vec{y}_t)$. This type of recall is free and happens immediately upon perception of $\vec{y}_t$. Note that associative recall embeds *recency* effects. To see this, simply let $R_a(\vec{y}_t) = \{\vec{y}_{t-s} : s \leq h\}$ so that the associative recall set contains the last h realizations of the state.

System 2: Costly [Conscious] Recall. In addition to free associative recall, the agent can also consciously recall memories at a cost. We assume that the cost associated with this types of recall is proportional to Shannon's conditional mutual information function. Note that this cost is already conditional on the associative recall set because associative recall precedes any conscious attempt in recalling other types of memory. So the cost of recall is $\omega \mathbb{I}(X^t; S_t | R_a(\vec{y}_t))$, where

$$X^t \equiv (\vec{x}_{t-h})_{h \geq 0}. \tag{2.2}$$

Note that if $S_t = R_a(\vec{y}_t)$ then the cost is zero. The more the agent tries to remember, the more cost he has to pay. We assume that S_t is chosen as a subset of the agent's past memories $\mathcal{S}_{X,t}(Y^t)$ defined as:

Definition 1. Let $\bar{\mathcal{S}}_{X,t}$ be the set of all possible signals over X. Then,

$$\mathcal{S}_{X,t}(Y^t) \equiv \{s \in \bar{\mathcal{S}}_{X,t} | \mathbb{I}(X^t, s | Y^t) = 0\}. \tag{2.3}$$

The assumption guarantees that the agent has access to a rich set of signals, as long as the recalled signals are not more informative of X^t than the information that is contained in the agent's long-term memory.

Agent's Problem. Given the primitives of the problem at time t, the agent solves:

$$\max_{R_{c,t} \subset \mathcal{S}_{X,t}(Y^t)} \mathbb{E}\left[\max_{\vec{a}_t} \mathbb{E}\left[v(\vec{a}_t, \vec{x}_t) | \bar{R}_t\right] - \omega \mathbb{I}(X^t, R_{c,t} | R_a(\vec{y}_t)) | R_a(\vec{y}_t)\right]$$

$$s.t. \quad \underbrace{\bar{R}_t}_{\text{total recall}} = \underbrace{R_a(\vec{y}_t)}_{\text{context}} \cup \underbrace{R_{c,t}}_{\text{conscious recall}} \tag{total recall}$$

3 Memory and Overpersistence

In this section, we apply the general framework developed in the previous section to the problem in which individuals make forecasts given past observations of a known stochastic process, analogous to the experimental setting in LMT. In this framework, individuals observe realizations of an AR(1) process:

$$y_t = \rho(y_{t-1} - \bar{y}) + \bar{y} + \epsilon_t, \tag{3.1}$$

and make predictions about the future realization of the stochastic process, $\hat{y}_{t+h}$.

Critically, the agent faces uncertainty over the true long-run mean of the process, $\bar{y}$. If he has access to the full database of past observations, the mean of past observations will converge to the true mean $\bar{y}$: in other words, the true long-run mean is measurable given the full memory bank M^t. However, this information is costly for the individual to access from his memory, and thus he can only obtain a noisy signal from his memory subject to informational costs.[1]

System I vs System II: The associative recall set at time t induces a belief about the long-run mean $\bar{y}$:

$$\bar{y}^I | I_t \sim N(\bar{y}^I, (\tau_I)^{-1}). \tag{3.2}$$

We shall assume in our setting that the System I mean of the long-run mean is given by y_t. Intuitively, this is reflecting our assumption that the individual automatically has access to the most recent observation of the stochastic process.

The agent then optimizes over the posterior-variance the following prediction task (subject to information costs):

$$\arg \min_{var(\bar{y}|I_t,a_t)} \gamma E[(\hat{y}_{t+h} - y_{t+h})^2] + \omega \log \left(\frac{var(\bar{y}|I_t)}{var(\bar{y}|I_t, a_t)} \right). \tag{3.3}$$

As expressed in the above equation, the individual is seeking to minimize prediction error over future outcomes y_{t+h}, not directly over the long-run mean $\bar{y}$. This not only better reflects the incentivization schemes in experimental data, but also highlights the insight that recall of long-run mean is more important for processes of low persistence.

Specifically, note:

$$\partial E[(\hat{y}_{t+h} - y_{t+h})^2] / \partial var(\bar{y}|I_t, a_t) = (1 - \rho^h)^2. \tag{3.4}$$

[1] We shall assume that the true auto-regressive coefficient ρ is known perfectly to the individual. We do not think it is necessary to take a stance on whether the individual has a correct persistence parameter: for now, we will assume that the agent does have a correct (and dogmatic) belief about the persistence, to emphasize the role that the imperfect recall of the long-run mean plays in explaining the overpersistence puzzle.

In other words, the importance of learning about the long-run mean is greater for less persistent processes (as the most recent data-point is less informative of future realizations), and increasing in the horizon of prediction.

Solving for the FOC implies that the final variance about the long-run mean is given by:

$$var(\bar{y}|I_t, a_t) = \min\left\{\frac{\omega}{\gamma(1-\rho^h)^2}, \tau_I^{-1}\right\} = \tau_{II}^{-1}, \tag{3.5}$$

which implies

$$\hat{y}_{t+h} = \rho^h y_t + (1-\rho^h)\left(\frac{\tau_I}{\tau_{II}}\bar{y}^I + \left(1 - \frac{\tau_I}{\tau_{II}}\bar{y}\right)\right) + \epsilon_t. \tag{3.6}$$

Summarizing, we obtain the following proposition:

Proposition 1. 1. The individual chooses to deploy costly memory retrieval iff $1 > \frac{\omega}{\gamma}\frac{\tau_I}{(1-\rho^h)^2}$. This happens if:

- Retrieval is sufficiently cheap (low ω),
- Precision is sufficiently important (high γ),
- (Subjective) precision of System I recall is sufficiently low (τ_I low),
- Persistence is sufficiently low (low ρ),
- Horizon of prediction h is sufficiently high.

2. Normalizing $\bar{y} = 0$, and assuming $\bar{y}^I = y_t$ (the automatic recall set only contains the most recent observation) we obtain the following expression for the average h period forecast:

$$\hat{y}_{t+h} = \min\left\{1, \left(\rho^h + \frac{\omega}{\gamma}\frac{\tau_I}{1-\rho^h}\right)\right\} y_t \tag{3.7}$$

Let us now explore the implication of our model in explaining individual forecast data.

3.1 Model predictions

Explaining the overpersistence puzzle:

The formula above implies that our general model speaks very naturally to the problem of overpersistence: when one regresses the agent's forecasts on the most recent realized observation, the implied autocorrelation coefficient is exaggerated by $\frac{\omega}{\gamma}\frac{\tau_I}{1-\rho^h}$. Intuitively, the most recent observation not only has a mechanical predictive power over future outcomes (given by ρ^h), but also partially inform the agent's beliefs of the long-run mean.

Sharply decaying predictability of past observations:

Next, our assumption that the individual relies on a combination of his readily accessible data (given by System I, $\{y_t\}$) and a long-term memory retrieval implies that the predictability of forecasts based on recent outcomes sharply decays for past data points.

Specifically, suppose we run the following regression:

$$\hat{y}_{t+h} = \sum_{k=0}^{H} \beta_k y_{t-k}. \tag{3.8}$$

While this sharp decay feature is also shared by a rational model with perfect memory, the rational model also predicts $\beta_0 = \rho^h$, which is not seen in the data.

On the other hand, suppose that the agent instead maintains a noisy memory state and costly storage, as is given by da Silveira and Woodford (2019). In their framework, the individual maintains a noisy memory state m_t that transitions stochastically based on the current memory state and the most recent observation y_t. For noisy memory models, we then obtain: $\beta_k = (1-\rho^h)\delta^k$, where δ^k governs the cost of moving the recent observations to long-term memory. Similarly, for costly perception models (with an otherwise costless memory retrieval), the weight the on past observations of y_t should be an exponential decay. In fact, if the agent is discounting the noisy signals in a Bayesian way, then $\beta_0 < \rho^h$ as well.

Thus, the weight of past observations (other than the most recent observation) on agent predictions should decay exponentially for these class of models.

3.2 Generalization to multiple horizon predictions

To better adapt our model to the empirical setting, which shall be the data collected by Landier, Ma and Thesmar (2020), we generalize our above model to allow for multiple horizon predictions. For simplicity, we shall assume that the agent now simultaneously predicts the next period observation, as well as a longer-horizon outcome, y_{t+h}, where $h > 1$. We shall assume that the agent is equally incentivized for accuracy for both horizons:

$$v(a_t, x_t) = -\frac{1}{2}\gamma \begin{bmatrix} y_{t+1} - \hat{y}_{t+1} \\ y_{t+h} - \hat{y}_{t+h} \end{bmatrix}^T \begin{bmatrix} y_{t+1} - \hat{y}_{t+1} \\ y_{t+h} - \hat{y}_{t+h} \end{bmatrix}, \tag{3.9}$$

where $a_t = (\hat{y}_{t+1}, \hat{y}_{t+h})$. Then, a similar optimization as in the previous section yields the following prediction:

Proposition 2. The average individual forecasts for period $t+1$ and $t+h$ are given by:

$$\begin{aligned}\hat{y}_{t+1} &= \min\left\{1, \rho + (1-\rho)\left(\frac{\omega}{\gamma}\frac{\tau_I}{((1-\rho)^2 + (1-\rho^h)^2)/2}\right)\right\} y_t, \\ \hat{y}_{t+h} &= \min\left\{1, \rho^h + (1-\rho^h)\left(\frac{\omega}{\gamma}\frac{\tau_I}{((1-\rho)^2 + (1-\rho^h)^2)/2}\right)\right\} y_t.\end{aligned} \tag{3.10}$$

Let us denote $\beta^i_{1,h}$ as the regression coefficient of period $t+i$ forecasts on the most recent observation, for $i = 1, h$:

$$\begin{aligned}\hat{y}_{t+1} &= \beta^1_{1,h} y_t, \\ \hat{y}_{t+h} &= \beta^h_{1,h} y_t.\end{aligned} \tag{3.11}$$

Then, we shall define $\hat{\rho}^i_{1,h} = \beta^{1/i}_{1,h}$ to be the implied persistence, for $i = 1, h$ (in other words, we convert from the observed regression coefficient to the implied persistence parameter by taking the $1/h$-th power).

4 Empirical Tests of the Model

In this section, we evaluate the performance of the model using the data collected by Landier, Ma and Thesmar (2020) (henceforth LMT), which collects forecasts of AR(1) processes with different levels of persistence in a simple experimental setting. We summarize the setting in Section 4.1, estimate the model using the data in Section 4.3, and present model fit in Section 4.4.

4.1 The Setting

LMT performs a simple, large-scale experiment where participants are asked to make forecasts of AR(1) processes:

$$x_{t+1} = \mu + \rho x_t + \epsilon_{t+1}. \tag{4.1}$$

Each participant is assigned to a condition with a given level of persistence ρ. At the beginning of the test, participants observe 40 past realizations of the process. They then make predictions of future realizations in each round of prediction, after which they observe an additional realization.

In the baseline test, participants predict x_{t+1} and x_{t+2}, and are randomly assigned into conditions with $\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. In additional tests, participants predict x_{t+1} and x_{t+5}, and are randomly assigned into conditions with $\rho \in \{0.2, 0.4, 0.6, 0.8, 1\}$. In all tests,

participants are incentivized to forecast the conditional mean. More than 2000 participants across various conditions were recruited from the US general population through Amazon's Mechanical Turk platform. An additional 200 participants were recruited from Electrical Engineering and Computer Science undergraduates at MIT. LMT show that the properties of the forecasts are the same whether the process is described as a "stable random process" or explicitly described as an AR(1) process in Equation 4.1 (for a subsample of randomly selected MIT students).

4.2 Key Properties in the Data

LMT analyze the properties of the forecasts, and test the subjective persistence implied by the forecasts as a function of the objective persistence ρ. The subjective persistence, $\hat{\rho}$, is obtained by regressing the forecast $F_t x_{t+1}$ on x_t (or analogously by regressing $F_t x_{t+h}$ on x_t and take the $1/h$th power of the regression coefficient). A key pattern they find is that $\hat{\rho}$ is greater than ρ, and the gap is particularly pronounced when the true persistence ρ is small. In other words, there appear to be a stronger degree of overreaction when ρ is small. The finding is consistent with results from Bordalo et al. (2019c) using survey data covering 20 macroeconomic and financial series. Although one cannot easily infer subjective persistence in the field data, given that econometricians do not know forecasters' information sets or the true data generating process, Bordalo et al. (2019c) also detect that overreaction seems more pronounced for series with lower persistence by testing the predictability of forecast errors using forecast revisions building on the insights of Coibion and Gorodnichenko (2015).

LMT also find that commonly used models cannot fit the variation of overreaction across series with different levels of persistence very well. Some models imply too much adjustment based on the objective persistence ρ, while others imply too little. They find that one reduced-form empirical model works well in fitting the data, but it does not directly derive from existing theories of expectations formation. We describe these models and their fit in the data in more detail below.

In the following, we show the subjective persistence generated by our model fits the patterns in the data surprisingly well.

4.3 Model Estimation

To fit our model, we further discipline our model by endogenizing τ_I, the precision in the System I beliefs. We assume:

$$\tau_I^{-1} = \frac{\sigma_\epsilon^2}{1-\rho^2}, \tag{4.2}$$

where σ_ϵ^2 is the variance of ϵ_t. In other words, we impose that the System I variance, given only a single data point y_t, is given by the stationary variance of the stochastic process.[2]

Given our endogenization of τ_I, our model effectively has only one degree of freedom, given by $\frac{\omega}{\gamma}$, which governs the information cost relative to the rewards obtained in the prediction task. This implies that our model has the same number of parameters as the other models that we shall compare our fit to.

Given a value of $\frac{\omega}{\gamma}$, we can thus generate the forecasts:

$$\begin{aligned}\hat{y}_{t+1} &= \min\left\{1, \rho + (1-\rho)\left(\frac{\omega}{\gamma}\frac{1}{\sigma_\epsilon^2}\frac{1-\rho^2}{((1-\rho)^2+(1-\rho^h)^2)/2}\right)\right\} y_t, \\ \hat{y}_{t+h} &= \min\left\{1, \rho^h + (1-\rho^h)\left(\frac{\omega}{\gamma}\frac{1}{\sigma_\epsilon^2}\frac{1-\rho^2}{((1-\rho)^2+(1-\rho^h)^2)/2}\right)\right\} y_t.\end{aligned} \tag{4.3}$$

4.4 Model Fit

In the following, we present results on model fit. We study the subjective persistence in the data, implied by our model, and implied by other models for each value of true persistence ρ. The other models include the following. LMT provide a more detailed summary.

1. Adaptive expectations (Cagan, 1956; Nerlove, 1958):

$$F_t x_{t+1} = \delta x_t + (1-\delta) F_{t-1} x_t. \tag{4.4}$$

   That is, expectations of the future outcome is a weighted average of the current outcome and the past forecast of the current outcome.

2. Traditional extrapolative expectations (Greenwood and Shleifer, 2014; Hirshleifer et al., 2015; Barberis et al., 2015):

$$F_t x_{t+1} = x_t + \phi(x_t - x_{t-1}). \tag{4.5}$$

   That is, expectations are influenced by the current outcome and the recent trend, and $\phi > 0$ captures the degree of extrapolation.

3. Noisy information/sticky expectations Bouchaud et al. (2019):

$$F_t x_{t+h} = (1-\lambda) E_t x_{t+h} + \lambda F_{t-1} x_{t+h} + \epsilon_{it,h}, \tag{4.6}$$

[2]This assumption can be given a Bayesian justification, where it can be interpreted as the posterior of $\bar{y}$ given y_t, where the prior over $\bar{y}$ converges to a flat prior.

which can capture expectations with noisy signals Woodford (2003) or noisy perception: $E_t x_{t+h}$ is the rational expectation, $\lambda \in [0,1]$ depends on the noisiness of the signal, and $\epsilon_{it,h}$ also comes from the noise in the signal.

4. Diagnostic expectations (Bordalo et al., 2018b):

$$F_t x_{t+h} = E_t x_{t+h} + \theta(E_t x_{t+h} - E_{t-1} x_{t+h}). \tag{4.7}$$

That is, the subjective expectation is the rational expectation plus overreaction to the surprise (measured as the change in rational expectations from the past period).

5. Constant gain learning (Malmendier and Nagel, 2016; Nagel and Xu, 2019):

$$F_t x_{t+h} = \widehat{E}_t^m x_{t+h} = \widehat{a_{h,t}} + \widehat{b_{h,t}} x_t, \tag{4.8}$$

where $\widehat{a_{h,t}}, \widehat{b_{h,t}}$ are obtained through a rolling regression using all data available until t, with exponentially decreasing weights ($w_t^s = \frac{1}{\kappa^{(t-s)}}$).

6. Forward-looking extrapolative expectations, a reduced-form model proposed by LMT:

$$F_t x_{t+h} = E_t x_{t+h} + \gamma(x_t - E_{t-1} x_t). \tag{4.9}$$

That is, forecasters extrapolate surprise represented by the latest shock.

The model parameters are estimated by minimizing the mean-squared error (MSE) between the 1-period forecast implied by the model for a given parameter and the forecast in the data. This is the same as how we estimate our model in the above. These parameters are shown in Table 1 of LMT. Panels A, B, C of Figure 2 analyze the 1-period ahead forecast ($F_t x_{t+1}$), the 2-period ahead forecast ($F_t x_{t+2}$), and the 5-period ahead forecast ($F_t x_{t+5}$), respectively.

Figure 2, Panel A, shows the results for $h = 1$: the solid line represents subjective persistence in the data, the red solid circles represent subjective persistence implied by our model, and the additional symbols represent subjective persistence implied by other models. Figure 2 Panels B and C present results for additional horizons $h = 2$ (Panel A) and $h = 5$ (Panel B), for the models with clear term structures (adaptive and traditional extrapolative models do not), based on the same parameters for each model.

We see that the subjective persistence implied by our model is very similar to that in the data. The subjective persistence implied by other models does not fit the data well, especially when ρ is low. Some models imply too little variation of the subjective persistence

with the objective persistence (e.g., adaptive expectations and traditional extrapolative expectations), while others imply too much variation of the subjective persistence with the objective (e.g., diagnostic expectations, constant gain learning, sticky expectations). The only other model that performs well is a reduced-form empirical model proposed by LMT, the forward-looking extrapolative model. It does not yet have a microfoundation, and its behavior is very similar to what is predicted by our micro-founded model.

Table 1 further evaluates model fit by calculating the MSE between $\hat{\rho}^h$ implied by the model and $\hat{\rho}^h$ in the data, and the MSE between $F_t x_{t+h}$ implied by the model and $F_t x_{t+h}$ in the data. We calculate the MSE for the 1-period, 2-period, and 5-period forecasts. In almost all cases, our model (the first row) has the smallest MSE.

Table 1: Model Fit

This table shows the MSE between $\hat{\rho}^h$ in the model in columns (1), (3), and (5), and the MSE between $F_t x_{t+h}$ implied by the model and $F_t x_{t+h}$ in the data in columns (2), (4), (6). Columns (1) and (2) report results for the 1-period forecast; columns (3) and (4) report results for the 2-period forecast; columns (5) and (6) report results for the 5-period forecast. The adaptive expectations model is: $F_t x_{t+1} = \delta x_t + (1-\delta) F_{t-1} x_t$. The traditional extrapolative expectations model is: $F_t x_{t+1} = x_t + \phi(x_t - x_{t-1})$. The sticky expectations model is: $F_t x_{t+h} = (1-\lambda)\rho^h x_t + \lambda F_{t-1} x_{t+h} + \epsilon_{it,h}$. The diagnostic expectations model is: $F_t x_{t+h} = E_t x_{t+h} + \theta(E_t x_{t+h} - E_{t-1} x_{t+h})$. The constant gain learning model is: $F_t x_{t+h} = \widehat{E}_t x_{t+h} = a_{t,h} + \sum_{k=0}^{k=n} b_{k,h,t} x_{t-k}$. The forward-looking extrapolation model is: $F_t x_{t+h} = E_t x_{t+h} + \gamma(x_t - E_{t-1} x_t)$.

Forecast horizon	$h=1$		$h=2$		$h=5$	
MSE Type	$\hat{\rho}^h$	Forecast	$\hat{\rho}^h$	Forecast	$\hat{\rho}^h$	Forecast
Current model	0.005	497.1	0.002	724.4	0.001	689.9
Adaptive	0.035	495.7	.	.	.	.
Extrapolative	0.064	527.3	.	.	.	.
Sticy	0.117	556.2	0.140	786.1	0.197	814.6
Diagnostic	0.069	521.2	0.115	758.0	0.177	803.3
Least square learning	0.111	560.6	0.079	817.4	0.094	856.7
Exponential decay	0.067	526.8	0.039	749.5	0.033	736.3
Hyperbolic decay	0.039	516.6	0.026	754.1	0.038	770.2
Forward-looking extrapolative	0.007	496.9	0.005	725.2	0.009	743.6

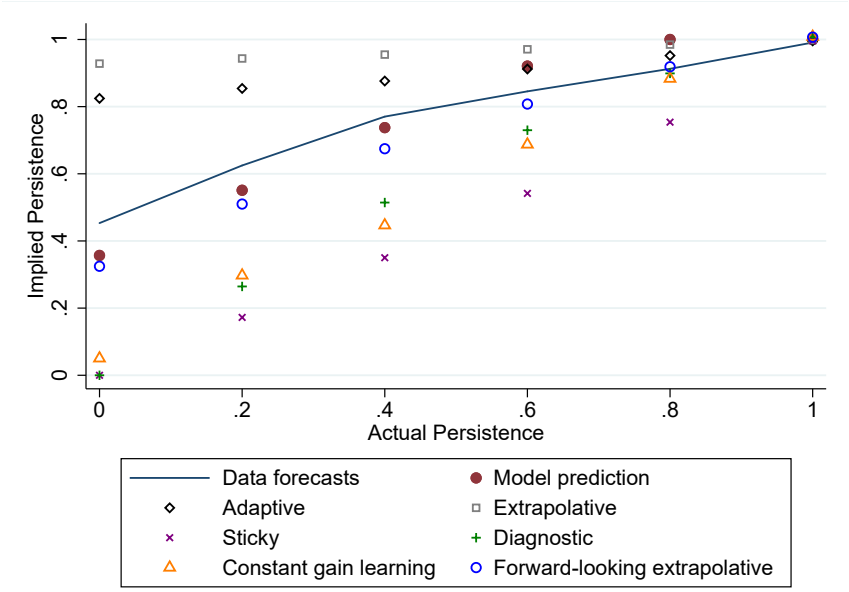
Finally, we provide a brief discussion of the intuition behind the better performance of our model. The alternative models can be categorized into two groups. For the first group, namely, adaptive expectations and traditional extrapolation, the models place a fixed weight on past observations that do not vary with the true persistence ρ. Consequently, with a given parameter, these models generate implied persistence that adapts too little to the situation (the curve is too flat).

For the second group, namely, diagnostic expectations and noisy information/sticky expectations, the models rely on rational expectations of the future forecasts. In particular, they converge to rational expectations when the true persistence is zero. The dependence on rational expectations and the adaptation turn out to be too strong in low persistence

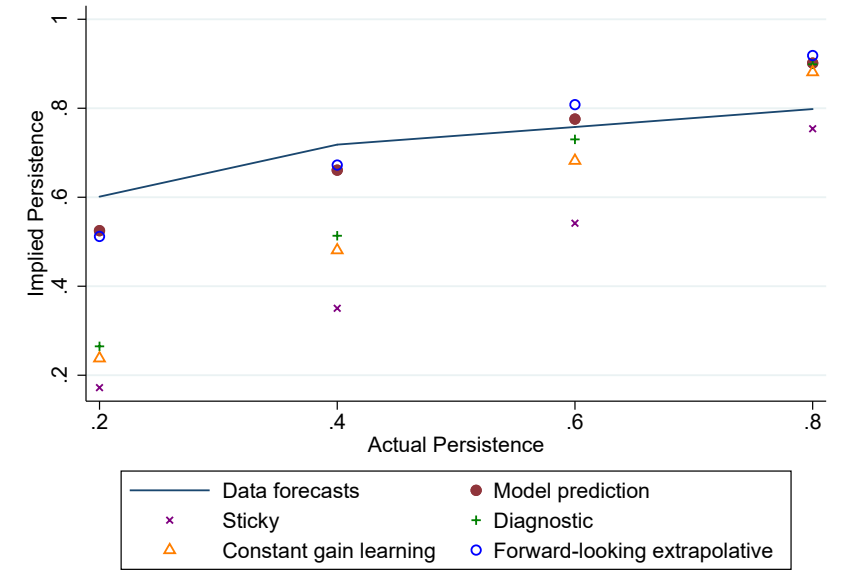
Figure 2: Subjective Persistence in the Data and in the Models

This figure shows the subjective persistence $\hat{\rho}^h$ as a function of the objective persistence ρ. The subjective persistence $\hat{\rho}^h$ is obtained by regressing $F_t x_{t+h}$ on x_t and taking the $1/h$th power of the coefficient. Panels A, B, C show results for $h = 1, 2, 5$ respectively. The solid lines represent $\hat{\rho}^h$ in the data. The solid red dot represents $\hat{\rho}^h$ implied by our model. The black hollow diamond represents $\hat{\rho}^h$ implied by adaptive expectations: $F_t x_{t+1} = \delta x_t + (1-\delta) F_{t-1} x_t$. The gray hollow square represents $\hat{\rho}^h$ implied by traditional extrapolative expectations: $F_t x_{t+1} = x_t + \phi(x_t - x_{t-1})$. The purple x represents $\hat{\rho}^h$ implied by sticky expectations: $F_t x_{t+h} = (1-\lambda)\rho^h x_t + \lambda F_{t-1} x_{t+h} + \epsilon_{it,h}$. The green cross represents $\hat{\rho}^h$ implied by diagnostic expectations: $F_t x_{t+h} = E_t x_{t+h} + \theta(E_t x_{t+h} - E_{t-1} x_{t+h})$. The orange triangle represents $\hat{\rho}^h$ implied by constant gain learning: $F_t x_{t+h} = \widehat{E_t^m} x_{t+h} = \widehat{a_{h,t}} + \widehat{b_{h,t}} x_t$. The blue hollow circle represents $\hat{\rho}^h$ implied by an empirical model with forward-looking extrapolation: $F_t x_{t+h} = E_t x_{t+h} + \gamma(x_t - E_{t-1} x_t)$.

Panel A. h=1

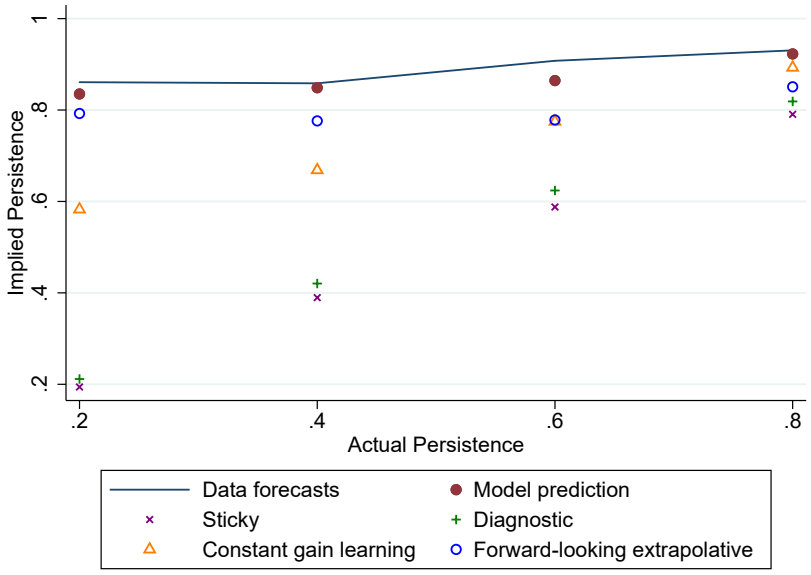


Panel B. h=2



Subjective Persistence in the Data and in the Models (Cont.)

Panel C. h=5



conditions (the implied persistence curve is too steep).

In our framework, due to limited memory, the forecaster conflates part of the transitory shock with changes in the long-run mean of the process. In this case, there is partial adaptation, and over-extrapolation also becomes more pronounced when the true process has low persistence. Moreover, to build connections with the reduced-form forward-looking extrapolative expectations model that LMT find to perform well in the data, one can consider the following simplified constant-gain learning process for the long-run mean:

$$\hat{\mu}_t = (1-G)\mu + Gx_t \implies F_t x_{t+1} = E_t x_{t+1} + (1-\rho)G(x_t - \mu). \tag{4.10}$$

This expression is reminiscent of the forward-looking extrapolative expectations model, where one replaces the long-run mean $\mu = E_{-\infty}x_t$ with the recent expectations $E_{t-1}x_t$. Consequently, our framework gives a way to justify the reduced-form formula in LMT (by microfounding the gain G in the simplified formula above) based on costly recall.

5 Conclusion

In this paper, we introduce a model of costly memory retrieval, where the agent balances the current news with memory of past observations. Our model matches key empirical features from recent research, namely, beliefs display stronger overreaction and over-persistence for

more transient processes and longer-horizon forecasts. The estimated model outperforms commonly used models in explaining these empirical patterns. A future direction is to connect our framework with a broader set of findings in memory research, such as recency effects and associative recall.

References

Abarbanell, Jeffry and Victor Bernard, "Tests of Analysts' Overreaction/Underreaction to Earnings Information as an Explanationfor Anomalous Stock Price Behavior," *Journal of Finance*, 1992.

Barberis, Nicholas, Robin Greenwood, Lawrence Jin, and Andrei Shleifer, "X-CAPM: An Extrapolative Capital Asset Pricing Model," *Journal of Financial Economics*, 2015, *115*, 1–24.

Barrero, Jose Maria, "The Micro and Macro Implications of Managers' Beliefs," Working Paper 2018.

Bondt, Werner FM De and Richard H Thaler, "Do security analysts overreact?," *American Economic Review*, 1990, pp. 52–57.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, Frederik Schwerter, and Andrei Shleifer, "Memory and representativeness," Working Paper 2019.

_ , Nicola Gennaioli, and Andrei Shleifer, "Diagnostic expectations and credit cycles," *Journal of Finance*, 2018, *73* (1), 199–227.

_ , _ , and _ , "Diagnostic expectations and credit cycles," *Journal of Finance*, 2018, *73* (1), 199–227.

_ , _ , and _ , "Memory, attention, and choice," *Quarterly Journal of Economics*, 2020, *Forthcoming*.

_ , _ , Rafael La Porta, and Andrei Shleifer, "Diagnostic expectations and stock returns," *Journal of Finance*, 2019, *74* (6), 2839–2874.

_ , _ , Yueran Ma, and Andrei Shleifer, "Over-Reaction in Macroeconomic Expectations," Working Paper 2019.

Cagan, Phillip, "The monetary dynamics of hyperinflation," *Studies in the Quantity Theory of Money*, 1956.

Coibion, Olivier and Yuriy Gorodnichenko, "Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts," *American Economic Review*, 2015, *105* (8), 2644–78.

da Silveira, Rava Azeredo and Michael Woodford, "Noisy Memory and Over-Reaction to News," Working Paper 2019.

Giglio, Stefano and Bryan Kelly, "Excess volatility: Beyond discount rates," *Quarterly Journal of Economics*, 2018, *133* (1), 71–127.

Greenwood, Robin and Andrei Shleifer, "Expectations of Returns and Expected Returns," *Review of Financial Studies*, 2014, *27* (3), 714–746.

Hirshleifer, David, Jun Li, and Jianfeng Yu, "Asset pricing in production economies with extrapolative expectations," *Journal of Monetary Economics*, 2015, *76*, 87–106.

Kahana, Michael Jacob, *Foundations of human memory*, OUP USA, 2012.

Kahneman, Daniel and Amos Tversky, "Subjective probability: A judgment of representativeness," *Cognitive Psychology*, 1972, *3* (3), 430–454.

Landier, Augustin, Yueran Ma, and David Thesmar, "Biases in Expectations: Experimental Evidence," 2020.

Ma, Yueran, Tiziano Ropele, David Sraer, and David Thesmar, "A Quantitative Analysis of Distortions in Managerial Forecasts," Working Paper 2020.

Malmendier, Ulrike and Stefan Nagel, "Learning from inflation experiences," *Quarterly Journal of Economics*, 2016, *131* (1), 53–87.

Nagel, Stefan and Zhengyang Xu, "Asset Pricing with Fading Memory," Working Paper 2019.

Nerlove, Marc, "Adaptive expectations and cobweb phenomena," *Quarterly Journal of Economics*, 1958, *72* (2), 227–240.

philippe Bouchaud, Jean, Philipp Krueger, Augustin Landier, and David Thesmar, "Sticky expectations and the profitability anomaly," *Journal of Finance*, 2019, *74* (2), 639–674.

Sims, Christopher A, "Implications of rational inattention," *Journal of Monetary Economics*, 2003, *50* (3), 665–690.

Wachter, Jessica A and Michael Jacob Kahana, "A retrieved-context theory of financial decisions," Working Paper 2019.

Wang, Chen, "Under- and Over-Reaction in Yield Curve Expectations," Working Paper 2019.

Woodford, Michael, "Imperfect Common Knowledge and the Effects of Monetary Policy," *Knowledge, Information, and Expectations in Modern Macroeconomics*, 2003.