

A Memory Based Model of Bounded Rationality

Sendhil Mullainathan*

April 9, 1998

Abstract

I investigate the impact of memory limitations on economic behavior with a model that is motivated by research in biology and psychology. Memory (along with access to some recorded information) is used by an agent who applies Bayes' rule to the recalled history as if it were the true one. The resulting behavior matches previously found experimental evidence on decision making biases. Applications to consumption and asset pricing produce results consistent with "anomalous" empirical findings: (i) consumption changes are negatively correlated with lagged income changes at the micro-level, but positively correlated at the macro-level; (ii) the marginal propensities to consume permanent income differs across income streams; (iii) stock prices exhibit excess volatility; (iv) price-earnings ratios forecast future stock returns; and (v) conditional on a single earnings surprise, prices under-react but conditional on a sequence of earnings surprises, all in the same direction, prices over-react. The model also predicts novel empirical relationships by tying the magnitude of these phenomena to measurable aspects of the labor income or firm earnings process.

*Economics Department, Harvard University. e-mail: mullain@fas.harvard.edu. I am indebted to my thesis committee, Drew Fudenberg, Larry Katz and Andrei Shleifer for their generous advice and encouragement, and to Marianne Bertrand, Ed Glaeser, Dave Laibson, and Dick Thaler for many helpful discussions. I have also greatly benefitted from comments by George Baker, John Campbell, Caroline Hoxby, Erzo F.P. Luttmer, Jay Patel, and participants at the Harvard University Labor and Finance lunches, Critical Theory and Labor seminars and the Harvard/MIT Behavioral Economics Seminar. Financial support from the Chiles Foundation is gratefully acknowledged.

1 Introduction

Can memory imperfections help explain observed economic behavior? Casual observation suggests that recollections do shape beliefs. An individual forecasting her income uses not only aggregate unemployment data, but also specific information drawn from more personal experiences. She may hear the anecdotes of a recently unemployed friend (whose skills closely mirror hers) or read about a foreign firm's plans to enter her industry. In fact, knowledge of these detailed events is exactly what separates an individual's forecast of her income from an econometrician's who must settle for only the statistical series.¹ Memory influences beliefs by shaping knowledge of these events, which by nature are qualitative and ephemeral. In this paper, I investigate whether the physiological limitations of human memory shed light on biases in the inference process.²

An advantage of focusing on memory is that scientific research on it is far more advanced than research on higher-order cognitive functions—such as problem solving techniques—and this permits a more grounded model. Accordingly, I use two stylized facts from biology and psychology research to model the memory technology. The first fact, termed *rehearsal*, states that remembering an event once makes it easier to remember that event again. Most students studying for an exam, by reading their lecture notes and repeatedly attempting to recall the material, take advantage of rehearsal. The second fact, termed *associativeness*, states that similarity of the memory to current events facilitates recall. Cues in today's events trigger memories that contain similar cues. Hearing your friend lament about how his Fiat has turned out to be a lemon may remind you of other Fiat horror stories. These two facts, which represent the core findings from memory research, form the basis of my model.

Two polar extremes mark the ways in which people may use this memory technology. They may be naive and apply Bayes' rule to the recalled history as if it were the true history. Alternatively, they may be sophisticated, by possessing complete knowledge of the recall technology and correct

¹The formal model developed in this paper allows individuals to access both recorded and recollected information.

²See Conlisk (1996) for a general survey of previous work on bounded rationality. Dow (1991) also presents a model of memory limitations, which examines optimal storage of information (in the context of search) given limited capacity.

for distortions in recall when forming forecasts. Each of these decision rules—as well as “partial adjustment” rules—has its appeal and undoubtedly a characterization of both is necessary. This paper takes the first step in this investigation by drawing out the implications of the naive model.³

Associativeness in recall generates a central property of beliefs: an event affects beliefs not only through the information it conveys but also through the memories it evokes. The friend’s lamentations about his Fiat influence perceptions not just because they provide one extra data point, but because they conjure memories that cast doubt on the reliability of Fiats. Recollections of a recent television expose may be jogged. This property implies that even *completely uninformative* signals can influence beliefs by altering the set of recalled memories. Even if one disregards these signals as noise, they have an indirect effect by altering perceptions of the past. Consider a political ad that contains a testimonial by a recently laid-off blue collar worker facing difficulties finding a decent job. Though the ad conveys little information, it may trigger other, more informative signals that suggest a recession. More generally, current events trigger memories with (on average) similar information. The added weight of these triggered memories leads to an over-reaction to news. For example, good news cultivates enthusiasm by producing a rosier view of the past. News draws forth reinforcing memories causing beliefs to respond “too much”.

Rehearsal generates another feature through the persistence of evoked memories. Even if the information in the original event has been discredited, the memories it triggered continue to be more memorable and hence continue to influence beliefs. A smear campaign can have lingering effects even after all the “facts” it proclaimed have been thoroughly debunked. The unflattering memories brought to mind stay, casting a negative shadow on the target. As another example, consider a judge instructing the jury to disregard the testimony they just heard. Even a well-intentioned jury would find it hard to fully comply with such a request. Even were they to ignore

³I chose to examine the naive rule first since experimental evidence suggests that individuals have neither accurate models of memory, nor correct for their memory mistakes in laboratory settings, making the naive model a natural first model to study. Of course, in cases with repetition and room for learning, sophistication may come to have more descriptive power. This makes it the next natural model to study and work in progress is examining it. This first pass also abstracts from recall effort. Individuals may work harder to remember certain events in the past over others. Such effort may take the form of mental exertion or the use of diaries to keep track of important information. Section 5 discusses these and other extensions to be pursued in future work.

the actual information, they must correct for the memories that it evoked as well. More generally, individuals under-react to news that invalidates or revises old information.

These results—and others derived through similar reasoning—match many of the experimentally found biases in human inference. For example, it has often been found that more salient information—a famous actress describing her battle with breast cancer as opposed to statistics—can have larger effects. The model generates this salience effect in two ways: (i) salient information may be more memorable and (ii) may be more effective at triggering supportive memories. As another example, several experiments have demonstrate the curse of knowledge: individuals find it hard to discount their private information in forming forecasts of how others will behave. I argue that one reason for this curse is that even if individuals disregard the direct effect of private information, the memories evoked by it still influence beliefs. In Section 3, I discuss the experimental evidence for these and several other psychological phenomena that can be interpreted in the current model. The model further relates the extent of such phenomena to a parameter of the stochastic process individuals are forecasting. This parameter measures (roughly) the ratio of transitory to permanent shocks. When it is large, the present value is a very noisy indicator of the permanent value, and inferences must rely heavily on past values. A greater reliance on history in turn means that memory distortions will be magnified. Thus increases in this parameter should magnify the behavioral biases. I argue in Section 3.9 that this relationship generates meaningful out of sample tests because in most economic applications, the relevant parameter can be measured using standard data sets.

This model also has implications for a wide range of economic behavior. In this paper, I sketch some implications for consumption and asset pricing. In a standard Permanent Income (PIH) model of consumption, memory distortions generate violations of the orthogonality predictions standard in such models: consumption changes *can* be predicted using lagged information. Moreover, under certain assumptions, the pattern of predictability resembles that found in existing empirical work. At the micro level, changes in consumption are negatively correlated with lagged income changes, while at the macro level they are positively correlated. While these two predictions have been

tested, the model provides a further prediction that has not: the correlation at the micro level should be more severe (roughly) as the ratio of transitory to permanent income shocks increases. The model also predicts that when there are multiple income sources, the marginal propensity to consume permanent income changes will be different for each stream. Moreover, the larger the MPC out of an income stream, the more negative the correlation between consumption and lagged income changes in that stream. Section 4.1 discusses these predictions, the intuition behind them, and related empirical evidence more carefully.

The second application focuses on asset pricing. I find that that while in the long run prices equal the net present value of the dividend stream, in the short run, memory distortions can generate mispricings. As noted before, beliefs respond to non-information or over-respond to intangible information. This produces one explanation for why prices seem to move with little apparent reason, or exhibit excess volatility. While this captures the response to intangible information, the response to more tangible information—earnings announcements—is more subtle. Under certain conditions, prices under-respond to a single earnings surprise, while over-responding to a sequence of earnings surprises all in the same direction. Following a single unexpectedly positive earnings shock, future returns will be predictably high. But, following a sequence of positive earnings shocks, future returns will be predictably low. As these results suggest, the model does not in general produce mean reversion since a price change may reflect either an over or under-reaction to news. It does, however, produce conditional mean reversion: price earnings ratios predict future returns. Finally, one can relate the extent of these phenomena to the earnings process of the firm. For example, firms about which there is less information such as newer firms will experience greater biases. Section 4.2 discusses these predictions, the intuition behind them, and related empirical evidence more carefully.

The limitations of human memory seem to provide a fruitful approach to understanding bounded rationality. Several psychological phenomena can be unified into a single model. Economic applications to consumption and asset pricing generate results that explain existing empirical anomalies. The model also generates new predictions that can be tested on standard data sets. Models based

on bounded rationality often invoke fears of *post hoc* rationalization, fears that with a sufficiently flexible set of assumptions almost any behavior can be “explained”. The out of sample predictions are a first step in alleviating such fears. As a whole, the findings suggest that models incorporating realistic limitations on recall have strong, testable implications about economic behavior.

2 Setup

The basic framework examines an individual who forms expectations about a state variable. I will take this variable to be synonymous with permanent income in future discussions, but it can be many other things; a firms’ earning power, macroeconomic conditions, or an employee’s abilities are just a few examples. Forecasts of income clearly influence many decisions—savings, job search, or portfolio choice—and in Section 4.1, I explicitly study the consumption decision. Labor income moves for a variety of reasons, such as macroeconomic shocks, technological innovations, or changes in expectations about an individual’s ability. As these examples indicate, forming forecasts requires combining a diverse set of information. Some of this information is, loosely speaking, “hard” or readily available in records: income in prior months, unemployment or GDP. Other information is “soft” or harder to capture in records: a friend in a similar position being fired or a boss telling you that you are one of the best employees he has seen. This disjunction between hard and soft will be useful for the model that follows. Knowledge of soft information depends on memory, while knowledge of hard information typically does not. The remainder of this section formalizes the setup.

2.1 Environment

Let y_t be income at time t , which follows the stochastic process:

$$y_t = \sum_{k=1}^t \nu_k + \epsilon_t \tag{1}$$

where ϵ_t is a transitory shock distributed $N(0, \sigma_\epsilon^2)$ and ν_k is a permanent shock, whose structure I will describe shortly. y_t is observed by the individual and represents the hard information.

Each period with probability p an event e_t occurs, which will be the “soft” information. Each event e_t has two components: an informative component, x_t , and an uninformative one, n_t . An example of an event is hearing about a friend describing his recent unemployment experience. The length of his unemployment spell would be informative, while the fact that he has a pregnant wife with medical bills piling up would be uninformative.⁴ The model includes uninformative, or neutral components, because they will affect recall probabilities.

When there is no event, I will write $e_t = \emptyset$ and, abusing notation, will generally take $x_t = n_t = 0$. Conditional on an event occurring, they are distributed:

$$e_t = (x_t, n_t) \sim F(x_t, n_t)$$

where $E[x_t] = E[n_t] = 0$. The covariance, $\sigma_{xn} = cov(x, n)$, measures whether the neutral component typically appears with positive or negative information. In the above example, the neutral component would measure the dollar value of the wife’s medical bills. Since these will be high when the unemployment spell is long and the signal is indicative of a negative income shock, the covariance σ_{xn} would be negative. The permanent shock at time t will be defined as:

$$\nu_t = x_t + z_t$$

where $z_t \sim N(0, \sigma_z^2)$.⁵ Thus, while x_t tells her something about the shock that period, its informativeness is incomplete and depends on σ_x^2/σ_z^2 .

The process in equation (1) generates a signal extraction problem: the individual must separate out the permanent shocks to y_t from the transitory ones. A 5% income drop may represent a negative shock to permanent income or may only affect current income. Knowledge of both past events and y_k help to solve this inference problem. Events e_k are useful because they allow one to extract a component of the time k innovation (x_k) that is *for sure* permanent. Past income

⁴Of course, as the example also illustrates, every part of an event will have *some* information content, and the dichotomy between x_t and n_t merely simplifies this spectrum.

⁵A slight awkwardness in this definition should be noted. The variance of z_t is higher when there is a shock than when there is none. In this sense, “signal” may not be a completely accurate word. This assumption does not drive the results; I use it so that the residual variance of z_t conditional on observing x_t is constant, allowing a more transparent analysis.

realizations y_k are useful because they allow one tease out the remainder of the permanent shock (z_k) but with less certainty. Repeated observations of high income will suggest a permanent rise.⁶

Finally, I describe the timing of the game. At the beginning of each period, she sees that period's events (if any). She then combines this knowledge with past information to form a forecast. At the end of the period the true value of y_t is observed. The game is then repeated.

2.2 Memory

2.2.1 Formal setup

Memory will be modeled as a stochastic map that transforms true history into perceived history.⁷

Let history, h_t , be a vector that includes y_k and e_k for $k < t$:

$$h_t = (y_1, \dots, y_{t-1}, e_1, \dots, e_{t-1})$$

Memory maps h_t into a random variable h_t^R . I begin by making mathematical assumptions about the nature of this map and then use experimental evidence to characterize the remainder.

As discussed, past values of income are hard information readily available in records. I, therefore, assume that y_k will be recalled perfectly. Events, on the other hand, characterize soft information, and are more prone to be forgotten. Formally, write recalled history as $h_t^R = (e_1^R, e_2^R, \dots, e_{t-1}^R, y_1, \dots, y_{t-1})$. Notice that in the recalled history, y_k is unaffected, whereas e_t is transformed into a random variable, e_t^R whose value is governed by:

$$e_k^R = \begin{cases} e_k & \text{with probability } r_{kt} \\ \emptyset & \text{with probability } 1 - r_{kt} \end{cases}$$

The probability that event e_k is recalled at time t is denoted by r_{kt} , where these probabilities are applied independently across events, though algebraically the probabilities may be linked.⁸ When

⁶Contrast with the case where y_t follows a standard random walk. Then, y_{t-1} is the only information in the past needed to forecast y_t . For the model formulated in this paper, \hat{y}_{t-1} is a sufficient statistic for all past information. This is an artifact, however, of the simplicity of the model. If we complicate it by assuming that different events have different levels of mean reversion rather than all being permanent, this will no longer be true. Forecasts must then rely directly on all past y_k and e_k .

⁷The recall process readily lends itself to a probabilistic interpretation. Casual conversation consists of phrases such as “more likely to remember” and experimental work supports this. James(1890) seems to present the first probabilistic interpretation of memory, though of course he does not use this terminology.

⁸Formally, conditional on r_{kt} and $r_{j\tau}$, R_{kt} and $R_{j\tau}$ are independent.

an event is forgotten, it is exactly as if no event occurred that period. A metaphor may help. Picture history as a series of boxes, one for every time period. Each box contains the details of that period's event. An empty box signifies that no event occurred that period. Memory goes to each box and flips a coin with weight r_{kt} to determine if the event in that box will be remembered; if forgotten, the box appears empty to the individual.

Notice some of the implicit assumptions made in this specification. Individuals do not remember distorted versions of events: they either remember them or not. They also do not "remember" events that never happened. Finally, a forgotten event matches a non-event, so that there is no feeling of "I think something happened but I'm not sure what". Weakening of these assumptions might all be useful tasks for the future. To complete the model, I need to specify r_{kt} . I turn to the scientific evidence for motivation on how to specify this.

2.2.2 Evidence on Memory

Research by biologists and psychologists has generated a consensus about two essential features of memory.⁹ The first, *rehearsal*, states that recalling a memory increases future recall probabilities. Students quickly recognize this property: repetition strengthens memories. The second, *associativeness*, states that events more similar to current events are easier to recall. For example, hearing a friend talk about his vacation will invoke memories of one's own vacations. Associativeness may arise because events serve as cues that help "find" lost memories.¹⁰ The evolutionary advantage of these two properties are easy to understand. Frequently encountered phenomena and memories similar to current circumstances are both more relevant.¹¹ As well as being intuitive, these two properties are also supported by an array of experimental evidence.

In describing the experimental evidence for rehearsal, I begin with the very small and work

⁹Schacter (1996) presents an excellent overview of this literature, one that I draw upon.

¹⁰An analogy helps to clarify this point. Remembering an event is like finding a house. The cues provided associations that help get you in the right block.

¹¹I have not formally pursued such intuitive notion to get at a more evolutionary or optimizing basis of memory. An accurate model along these lines seems well beyond our current abilities since it would require a precise understanding of the constraints on what memory mechanisms are even feasible. Theorizing that omits such constraints runs the risk of telling empty "just so" stories.

my way up. Biological evidence suggests that memories are “stored” in the synaptic connections between neurons. Repeated firings across these increases connection strength—both by increasing their number and their importance—and enhances the memory. In the case of the gill withdrawal reflex of the marine snail *Aplysia*, we can map this process. When first presented with a noxious stimulus (such as an electric shock), the creature withdraws its gill. As the number of such stimuli increases it responds faster, more frequently, and the persistence of the response across experiments increases. Dissection reveals an increase in the neural connections between areas controlling gill withdrawal and areas experiencing the stimulus. Consistent with rehearsal, the strength of these connections increases with repetition.¹² Rehearsal has also been found in experiments with humans, where two groups of subjects memorize the same list of words. One group then practices recall of this list periodically, while the other does not.¹³ After the same time has elapsed for both groups, they are asked to recall the list. Typically, one finds that the group that has been periodically recalling the list shows higher recall. This phenomena can also be observed in more natural settings. For example, Spanish learned in grade school is more likely to be retained if used than if unused. That these findings should seem so obvious is a testament to the intuitive appeal of the rehearsal assumption.

A good example of how associativeness can operate is found in experiments on classical conditioning.¹⁴ In the seminal experiment, Pavlov trained dogs by ringing a bell and then feeding them. Once this procedure was repeated several times (notice the role of rehearsal), the ringing of the bell alone would induce salivation: the dogs had come to associate the ringing of the bell with the imminent meal. Also, notice in the *Aplysia* gill withdrawal reflex experiments described above, that the

¹²See Kandel, Schwartz and Jessell (1991). A contrasting effect is habituation, wherein synaptic strength decreases with frequency. This corresponds to the idea that novel stimulus receives notice which lessens as the novelty wears off. I ignore this property because it a property of *attention* rather than of memory. The idea that a repeated stimulus results in longer term responses also underlies classical conditioning, a point I return to shortly.

¹³Both groups look at the true word list only once.

¹⁴See MacKintosh (1983) summarizes conditioning. Laibson (1997) derives a theory of consumption based on preferences that exhibit a form of conditioning. Our papers differ because I focus on expectations rather than preferences. The similarity is interesting, however, and suggests that a memory model, in which individuals must use past experience to forecast preferences, potentially provides one microfoundation to the preferences used by Laibson (1997).

neural pathways between the areas controlling gill withdrawal and the areas sensing the stimulus strengthen, an example of associativeness. The importance of associativeness in every day recall has been emphasized by Tulving and his colleagues, who study the role of cues in recall. Subjects learn a list of words in which each target word is paired with a cue word. The subjects are then asked to remember the target words, and are either provided with the associated cue or not. A broad set of such experiments finds that recall of the target words is higher when the paired word is present.¹⁵ A related example of this phenomena is subjects who learn the sentence (Anderson et. al., 1976)

The fish attacked the swimmer

They are more likely to remember this sentence if given the cue “shark” than if given no cue at all. Notice, however, that “shark” never appears in the sentence, which illustrates that associativeness likely operates also through conceptual similarity. Yet another example is provided by the importance of physical environment. In one experiment, subjects either learned word lists under water or on land, and were asked to recall them either under water or on land. Recall was higher under water for those who learned the list under water and on land for those who learned the list on land. Mood has a similar effect: things learned in one mood (e.g. depressed) are better remembered in that mood.

As these studies demonstrate, both rehearsal and associativeness have a strong experimental basis. In fact, the most popular models of memory (and neural function generally), Parallel Distributed Processing Models, possess both features.¹⁶ Nevertheless, I do not mean to imply that these are the only important facts about memory. Let me cite the two most interesting omissions. First, researchers now believe that certain memories are episodic (the time you tasted caviar), while others are semantic (you dislike caviar). This distinction is interesting because semantic memories

¹⁵These paired words sometimes share a natural connection, such as “brain” and “mind” or “brain” and “drain”, and sometimes are unrelated, such as “brain” and “doughnut”. The findings hold in both cases though the magnitude of the effect varies.

¹⁶See Rumelhart and McClelland (1986). The other key feature of PDP models is their massive parallelism. Recall (and more generally computation) in these models is not the outcome of a single “smart” unit. Instead it arises from the interaction of several “dumb” units.

may not possess all the episodes that gave rise to them. That these two different memory systems exist may have interesting implications for the dynamics of learning and the evolution of heuristics. Second, memory seems to be reconstructive in nature. As Neisser (1967) puts it, memory is like a paleontologist putting together a dinosaur from bits of old bone.¹⁷ The process of reconstruction uses *a priori* theories to put together the pieces, so facts that deviate from these theories will more likely be forgotten. In a seminal experiment, Bartlett (1932) demonstrated how in recalling stories, subjects often edit out inconsistent parts. These omitted findings may have interesting implications for economics, and may be the object of future work. I ignore them for the time being, however, because they lack the mass of evidence that supports the other two assumptions and because they are analytically more vague.

2.2.3 Formalism

Three parameters appear in the formalism: \underline{m} (the baseline recall probability), ρ (which quantifies rehearsal), and χ (which quantifies associativeness). Assume that all are between zero and one and that $\underline{m} + \rho + \chi < 1$. Let R_{kt} denote the random variable which equals 1 iff event k is recalled at time t .¹⁸ Note that $E[R_{kt}] = r_{kt}$. With this notation, we can write:

$$r_{kt} = \underline{m} + \rho R_{k(t-1)} + \chi a_{kt} \tag{2}$$

The first term equals the baseline recall probability for all memories, \underline{m} . The second term captures rehearsal. Events recalled in the last period get a “boost” of ρ . This formalism of rehearsal may seem awkward. Consider two events: e_{t-2} which occurred two days ago and e_{t-20} which occurred twenty days ago, and suppose neither is remembered yesterday. Then (holding the third term constant) both have the same recall probability. Recall appears to display sharp, rather than smooth, decay. This awkwardness is superficial. Proposition 1 demonstrates that *in expectation*, recall probabilities do exhibit exponential decay. Alternatively, I could build exponential decay directly into the dynamics of memorability, so that it occurs not only in expectation but on every

¹⁷This is drawn from Schacter (1996).

¹⁸Take $R_{(t-1)t} = 1$ and $r_{(t-1)t} = 1$ for the initial conditions.

realization and results would not change.

The third term captures associativeness where a_{kt} measures the similarity of event e_k to e_t . The events $e_k = (x_k, n_k)$ and $e_t = (x_t, n_t)$ are two points on a plane. Similarity can then be defined as a negative function of the distance between the points. Let $c : (\Leftrightarrow\infty, \infty) \rightarrow (0, 1)$ be a closeness function (that is, an inverse distance function). Then similarity is defined as:

$$a_{kt} = \frac{1}{2} (c(x_t \Leftrightarrow x_k) + c(n_t \Leftrightarrow n_k))$$

and with the assumption that $a_{kt} = 0$ if either e_k or e_t is a non-event. I will take the specific function $c(x) = e^{-x^2}$, which allows one to write: $a_{kt} = \frac{1}{2} (e^{-(x_t-x_k)^2} + e^{-(n_t-n_k)^2})$. Thus $0 < a_{kt} < 1$ and $a_{tt} = 1$.

Substituting back in to the original equation provides:

$$r_{kt} = \underline{m} + \rho R_{k(t-1)} + \chi \frac{1}{2} (c(x_t \Leftrightarrow x_k) + c(n_t \Leftrightarrow n_k)) \quad (3)$$

and recall that I assume that $\underline{m} + \rho + \chi < 1$. It will also be useful to define $f_{kt} = 1 \Leftrightarrow r_{kt}$ to be the probability of forgetting, and $F_{kt} = 1 \Leftrightarrow R_{kt}$ to be the indicator variable for having forgotten event k . This probability can be written as:

$$f_{kt} = \underline{f} + \rho F_{k(t-1)} \Leftrightarrow \chi \frac{1}{2} (c(x_t \Leftrightarrow x_k) + c(n_t \Leftrightarrow n_k)) \quad (4)$$

where $\underline{f} = 1 \Leftrightarrow \underline{m} \Leftrightarrow \rho$ is the baseline forgetting probability.

Finally, note that unlike the other basic assumptions of the model, the choice of functional form here is arbitrary. I could well have included an interaction term between associativeness and rehearsal or higher order terms. As another example, I might have allowed for limited capacity so that only a finite set of memories can be recalled at any time, which would generate crowd out. The intuition behind the results that follow does not rely on the functional form, though some of these extensions are clearly worth investigating.

3 Basic Results

3.1 Some Useful Definitions

Two definitions will help to characterize behavior:

Definition 1 *The **vividness** of an episode e_t equals*

$$\mathcal{V}(e_t) = E[a_{kt}|e_t]$$

An event is vivid if it is very similar to a randomly drawn event. Intuitively, if an event shares many cues with commonly encountered events, it will be more vivid.¹⁹ A related property will be evocativeness.

Definition 2 *The **evocativeness** of event e_t is:*

$$\mathcal{E}(e_t) = E[x_k a_{kt}|e_t]$$

An event's evocativeness measures the information content of the memories triggered by that event. If today's event is e_t , then a_{kt} measures the strength of its association with event (memory) e_k , while x_k measures the information content of that past event. The expectation of this product, therefore, measures the average information content of memories brought forth by associativeness. To illustrate evocativeness, consider the event $e_t = (x_t, n_t) = (1, 1)$. The evocativeness of this event has two parts. First, $x_t = 1$ implies positive evocativeness. Other x_t close to 1 will be evoked, leading to an oversampling of positive memories and positive evocativeness. Second, $n_t = 1$ can have a positive or negative impact on evocativeness depending on σ_{xn} . Since $n_t = 1$, other events with positive n_k are triggered, but the information content of these events clearly depends on σ_{xn} . When σ_{xn} is zero, knowing that $n_k > 0$ says nothing about x_k , so that the effect on evocativeness is zero. If σ_{xn} is positive, knowing $n_k > 0$ tells us that $x_k > 0$: positive events are selectively triggered causing a positive effect on evocativeness. Finally, when σ_{xn} is negative, $n_k > 0$ tells us that $x_k < 0$ generating a negative effect on evocativeness. Summarizing, the marginal impact of a change in n_t on evocativeness is proportional to σ_{xn} .

¹⁹This has the unfortunate feature that very unusual events are labeled as less vivid. For more on this, see footnote 21.

3.2 Dynamics of Recall

Proposition 1 *Let $k < t$. Forgetting probabilities then satisfy:*

$$E[f_{kt}|e_k] = (\underline{f} \Leftrightarrow \chi \mathcal{V}(e_k)) \frac{1 \Leftrightarrow \rho^{t-k}}{1 \Leftrightarrow \rho} \quad (5)$$

$$\lim_{t \rightarrow \infty} E[f_{kt}|e_k] = \frac{\underline{f} \Leftrightarrow \chi \mathcal{V}(e_k)}{1 \Leftrightarrow \rho} \quad (6)$$

Proof: Notice that

$$E[f_{kt}|e_k] = \underline{f} \Leftrightarrow \chi \mathcal{V}(e_k) + \rho E[f_{k(t-1)}|e_k]$$

Recursive substitution yields:

$$(\underline{f} \Leftrightarrow \chi \mathcal{V}(e_k))(1 + \rho^2 + \dots + \rho^{t-k-1}) = (\underline{f} \Leftrightarrow \chi \mathcal{V}(e_k)) \frac{1 \Leftrightarrow \rho^{t-k}}{1 \Leftrightarrow \rho}$$

Taking limits produces the second part of the proposition. ■

This proposition provides two facts about this model of memory. First, recall probabilities decay exponentially over time: further back memories have higher chances of being forgotten. Experimental evidence on recall probabilities indicate that exponential decay of memories fits the data rather well.²⁰ Second, vividness raises memorability. More vivid memories share more cues with commonly encountered events. This makes it more likely that they will be triggered through associativeness. The increased recall probability will be useful when thinking about the notion of how “salient” information is, and why more salient information may have greater impact.²¹

3.3 Structure of Expectations

3.3.1 Perfect Memory Forecasts

The perfect memory forecast will serve as a useful base case against which one can compare the forgetful forecast. Recall that z_k and ϵ_k are normally distributed, and all the x_k relevant for forecasting y_t are known with certainty. This implies that the residual uncertainty is normal.

²⁰See Crovitz and Schiffman (1974). A power function, however, seems to fit better.

²¹This result, however, has the unfortunate property that outliers, very unusual events, have *lower* recall probability, contrary to one’s intuition. One resolution to this problem may be found in allowing for the possibility that unusual events may receive greater attention, and that attention may increase memorability.

Assuming that prior distributions are normal, what remains is a normal learning model. Let \hat{y}_0 and $\hat{\sigma}_0^2$ be the prior distribution over ν_0 . Let the (normal) posterior distribution over y_t be summarized by the mean $\hat{y}_t(h_t, e_t)$ and variance $\hat{\sigma}_t^2(h_t, e_t)$ where I'll typically drop the dependence on h_t and e_t for simplicity.

In steady state beliefs satisfy (see Lemma 1 in the appendix):

$$\hat{y}_t(h_t, e_t) = x_t + \sum_{k=1}^{t-1} \left[\lambda^{t-k} x_k + (1 \Leftrightarrow \lambda^{t-k}) \Delta y_k \right] \quad (7)$$

$$\hat{\sigma}_t^2(h_t, e_t) = \sigma_*^2 \quad (8)$$

where $\Delta y_k = y_k \Leftrightarrow y_{k-1}$ and

$$\lambda = \frac{\sigma_\epsilon^2}{\sigma_*^2 + \sigma_\epsilon^2}$$

Here λ measures the signal-to-noise ratio in y_t .

Understanding the marginal impact of different variables will improve intuition about the forecast rule. First, x_k influences forecasts one-for-one. Its impact is the sum of two terms. There is a direct effect which contributes $\lambda^{t-k} x_k$ and an indirect effect from Δy_k , because $\Delta y_k = x_k + z_k + \epsilon_k \Leftrightarrow \epsilon_{k-1}$, that contributes $(1 \Leftrightarrow \lambda^{t-k}) x_k$. Summing shows that the total coefficient on x_k is unity. Second, Δy_k enters forecasts with weight $1 \Leftrightarrow \lambda^{t-k} < 1$ as is clear from the formula. Third, y_k influences forecasts at $\lambda^{t-k-1}(1 \Leftrightarrow \lambda) < 1$ because it enters in Δy_k and in Δy_{k+1} . Both y_k and Δy_k have a less than one-for-one impact because both are noisy estimates of permanent income (or permanent income changes). That x_k has greater impact reiterates the importance of events in separating signal from noise. They show the individual a portion of the income change that for sure is permanent. Fourth, n_k has zero impact as expected: neutral components convey no information. Finally, λ measures the importance of history in forecasts. As λ increases, older y_k receive greater weight.

3.3.2 The Limited Memory case

To calculate the posteriors of a forgetful individual, I make a key behavioral assumption. I assume that in making forecasts, the forgetful individual applies the forecasting rule in equation 7 to the

recalled history. In other words, she takes the the recalled history as the *true* history. Let $\hat{y}_t^R(h_t^R, e_t)$ denote the mean and $\hat{\sigma}_t^{2R}(h_t^R, e_t)$ denote the variance of a (naive) forgetful Bayesian's posteriors. This assumption can then be stated formally.

Assumption 1 *The forgetful forecast applies the optimal forecasting rule to the recalled history:*

$$\begin{aligned}\hat{y}_t^R(h_t^R, e_t) &= \hat{y}_t(h_t^R, e_t) \\ \hat{\sigma}_t^{2R}(h_t^R, e_t) &= \hat{\sigma}_t^2(h_t^R, e_t)\end{aligned}$$

I have referred to this as the *naive* decision maker, in contrast to the *sophisticated* decision maker, who completely knows the model of memory and corrects his forecast rule accordingly.²² Deciding between these two polar models will be an important task, and one that requires a complete characterization of behavior in both cases.

I have chosen to investigate the naive case first because experimental evidence suggests that it describes behavior at least in the laboratory. Studies of individual's judgements of their own memories reveal inaccuracy in understanding their memory process.²³ Similarly, experiments have manipulated the memorability of information and tested whether individuals' decisions correct for this manipulation. Supportive of the naive assumption, decisions are insensitive to this manipulation.²⁴ Of course, repetition and room for learning, may dramatically alter these findings. Nonetheless, the findings suggest that characterizing the naive decision maker would be a useful first step.²⁵

Simple substitution gives the formula for the forgetful forecast:

$$\hat{y}_t^R(h_t^R, e_t) = x_t + \sum_{k=1}^{t-1} \left[R_{kt} \lambda^{t-k} x_k + (1 \Leftrightarrow \lambda^{t-k}) \Delta y_k \right] \quad (9)$$

$$\hat{\sigma}_t^{2R}(h_t^R, e_t) = \sigma_*^2 \quad (10)$$

²²Results in the sophisticated model depend on whether or not past forecasts are recalled.

²³See the literature on meta-memory as in Reder (1996).

²⁴See, for example, Trope (1978). This also resembles findings by Kahneman and Tversky (1982) on the availability heuristic, that individuals take more easily remembered events to also be more probable.

²⁵Preliminary results suggest that even more nuanced results may arise in the sophisticated model. For example, suppose that forecasts are not remembered but that zero-one decisions which condition on forecasts are remembered with certainty. Then, a herding problem akin to Banerjee (1992) arises. Consider an individual who remembers choosing 1 several times but currently faces information that suggests 0 is the best choice. For certain histories and parameter values, the weight of having chosen 1 in the past ("I must have had some reason to do it") will dominate and he will choose 1 again. But this implies that the 0 signal that he received this period will be jammed and he will be stuck in a herding equilibrium.

In words, forgetful forecasts look just like perfect recall forecasts except that forgotten events ($R_{kt} = 0$) are excluded. Note that \hat{y}_t^R is a random variable. Taking expectations over this random variable implies that events are weighted by their recall probability.

3.4 Forecast Errors and Bias

Define

$$err_t = y_t \Leftrightarrow \hat{y}_t$$

$$err_t^R = y_t \Leftrightarrow \hat{y}_t^R$$

$$err_t^m = \hat{y}_t \Leftrightarrow \hat{y}_t^R$$

to be the forecast error for the perfect recall forecast, the forecast error for the forgetful forecast and the *memory error* respectively. Note that:

$$err_t^R = err_t + err_t^m$$

so that the memory error measures how much of the forgetful forecast error is due to memory distortions. The following proposition establishes when the forgetful forecast is unbiased.

Proposition 2 *The forgetful forecast is unbiased if and only if average evocativeness is zero:*

$$E[err_t^R] = 0 \Leftrightarrow E[\mathcal{E}(e)] = 0$$

Similarly, conditioning on history the forecast error tends to zero in the long run if and only if average evocativeness is zero:

$$(\forall h_t) \lim_{j \rightarrow \infty} E[err_{t+j}^R | h_t] = 0 \Leftrightarrow E[\mathcal{E}(e)] = 0$$

Symmetry of the x_t and n_t distribution guarantees zero average evocativeness.

Proof: Break apart $err_t^R = err_t + err_t^m$ and notice that $E[err_t] = E[err_t | h_t] = 0$.

Therefore:

$$E[err_t^R] = E[err_t^m] = \sum_{k=1}^{t-1} \lambda^{t-k} E[f_{kt} x_k]$$

This equals $\Leftrightarrow \chi E[\mathcal{E}(e)] \frac{\lambda(1-\lambda^{t-1})}{1-\lambda}$ which equals zero if and only if $E[\mathcal{E}(e)] = 0$. For the second part, using the same logic:

$$E[err_{t+j}^R | h_t] = \lambda^j \sum_{k=1}^{t-1} \lambda^{t-k} x_k E[f_{k(t+j)} | h_t] + \sum_{k=t}^{t+j-1} \lambda^{t-k} E[f_{k(t+j)} x_k]$$

The first part of this equation tends to zero as $j \rightarrow \infty$ since $\lambda^j \rightarrow 0$. The second part tends to: $\Leftrightarrow \chi E[\mathcal{E}(e)] \frac{\lambda}{1-\lambda}$. Therefore, this equals zero if and only if average evocativeness is zero.

Finally, suppose that x and n are symmetrically distributed, that is $F(x, n) = F(\Leftrightarrow x, \Leftrightarrow n)$. Note that $2E[\mathcal{E}(e_t)]$ equals: $E[c(x_k \Leftrightarrow x_t) x_k] + E[c(n_k \Leftrightarrow n_t) x_k]$. The first part of this can be written as:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} c(x_k \Leftrightarrow x_t) x_k dF(x_k) dF(x_t) = \left(\int_0^{\infty} \int_0^{\infty} + \int_{-\infty}^0 \int_{-\infty}^0 + \int_0^{\infty} \int_{-\infty}^0 + \int_{-\infty}^0 \int_0^{\infty} \right) c(x_k \Leftrightarrow x_t) x_k dF(x_k) dF(x_t)$$

Perform the transformation x_k to $\Leftrightarrow x_k$ and x_t to $\Leftrightarrow x_t$ in the second and fourth integrals.

By symmetry of $dF(\cdot)$ and $c(\cdot)$, this allows one to write

$$\left(\int_0^{\infty} \int_0^{\infty} \Leftrightarrow \int_0^{\infty} \int_0^{\infty} + \int_0^{\infty} \int_{-\infty}^0 \Leftrightarrow \int_0^{\infty} \int_{-\infty}^0 \right) c(x_k \Leftrightarrow x_t) x_k dF(x_k) dF(x_t) = 0$$

A similar trick applies for $E[x_k c(n_k \Leftrightarrow n_t)]$. ■

Suppose average evocativeness is positive. Then, positive events are triggered more often than negative events. Memories, therefore, over-represent positive information and generate forecasts that are too large. Zero evocativeness guarantees that the set of sampled memories fairly represents (on average) both good and negative memories. The second part shows that even when one conditions on h_t biases eventually disappear. This demonstrates a ‘‘long-run correctness’’ property of the model.²⁶

Under symmetry of the distributions, the lack of bias is no surprise: nothing in the model distinguishes positive from negative. When the symmetry condition is weakened, however, biases

²⁶See Section 3.9 for a careful discussion of this long run learning and why it occurs.

can arise. An intuitive example can be had in the discrete case. Ignore the neutral component, and assume that x_t is one of $\{+g, \Leftrightarrow b, 0\}$ where $g, b > 0$ with probabilities, p_g, p_b and $1 \Leftrightarrow p_g \Leftrightarrow p_b$ respectively. Finally, to guarantee $E[x_t] = 0$, assume that $p_g g = p_b b$. In this case, the forecast errors will be biased if and only if $p_g \neq p_b$. Moreover, forecasts will be too high when $p_g > p_b$ and too low when $p_b > p_g$. The intuition is simple. Suppose $p_g > p_b$. Good events are more probable and more likely to be triggered by future associations. Their vividness will, therefore, be high as will their recall probabilities. But, of course, this will lead to a forecast that is too large. To abstract from such situations, I will impose a symmetry assumption.

Assumption 2 *Assume e_t is normally distributed (conditional on there being an event):*

$$e_t = (x_t, n_t) \sim N(0, \Sigma); \Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xn} \\ \sigma_{xn} & \sigma_n^2 \end{pmatrix}$$

3.5 Vividness, Evocativeness and Beliefs

With these definitions in hand, I now examine the determinants of beliefs.

Proposition 3 *The impact of event e_t on time t beliefs does not depend on its vividness, but does depend (positively) on its evocativeness. On the other hand, its impact on time $t + j$ beliefs depends on both vividness and evocativeness.*

Proof: From Lemma 3

$$E[\hat{y}_t^R | e_t] = x_t + \chi \mathcal{E}(e_t) \frac{\lambda(1 \Leftrightarrow \lambda^{t-1})}{1 \Leftrightarrow \lambda}$$

showing the dependence of evocativeness, and the absence of a vividness effect. The intuition here is simple. Evocativeness influences what memories are triggered and, therefore, has a direct effect on beliefs. Vividness only operates through increased memorability, which of course cannot have an impact on contemporaneous beliefs.

For the impact on future beliefs, Lemma 4 in the appendix shows that:

$$E[\hat{y}_{t+j}^R | e_t] = x_t \left(1 \Leftrightarrow \frac{f \Leftrightarrow \chi \mathcal{V}(e_t)}{1 \Leftrightarrow \rho} (1 \Leftrightarrow \rho^j) \lambda^j \right) + (\rho \lambda)^j \chi \mathcal{E}(e_t) \frac{\lambda}{1 \Leftrightarrow \lambda} (1 \Leftrightarrow \lambda^{t-1})$$

where we see as before the dependence on evocativeness. This is because the memories triggered at time t were rehearsed and, therefore, continue to have higher recall probability even at time $t + j$. Consistent with this, note that as $\rho \rightarrow 0$, the effect disappears. We also see here that vividness now plays a role. As we saw in Proposition 1, vividness increases memorability. Thus it increases the *marginal impact* of x_t by making it more likely to be recalled and used in forming beliefs. One implication is that when $x_t = 0$, changes in vividness have no impact: whether or not the event is recalled, it does not influence beliefs. ■

This proposition and its proof makes several points that are worth reiterating. Vividness plays no role in how an event influences beliefs at the time it occurs. It only matters as time passes and a chance to forget the event appears. By increasing memorability, vividness influences whether or not an event is remembered and thereby whether or not the information it conveys is used in the future.

Evocativeness measures the information content of the memories triggered by an event and, therefore, influences beliefs even at the time the event is occurring. An event with positive evocativeness, for example, disproportionately draws forth positive memories leading to a more positive forecast. Moreover, since these triggered memories persist (by rehearsal), evocativeness also influences future beliefs, though its effect diminishes over time (the ρ^j exponent). Summarizing, the current model decomposes the intuitive notion of “salience” into two components: vividness, which captures increased memorability, and evocativeness, which captures the ability of events to trigger supporting evidence. Both affect an event’s impact on beliefs but do so in different ways.

Finally, note that even if $x_t = 0$, an event can influence beliefs as long as evocativeness is positive. This is the heart of the next proposition, which deals with the impact of uninformative components.

Proposition 4 *Let $e_t = (0, n_t)$ be an uninformative event but with non-zero neutral component ($n_t \neq 0$). This event influences beliefs if and only if $\sigma_{xn} \neq 0$. The sign of this influence equals:*

$$\text{sign}(E[\hat{y}_t^R | e_t]) = \text{sign}(\sigma_{xn} n_t)$$

For a general event, changes in n_t influence beliefs if and only if $\sigma_{xn} \neq 0$.

Proof: Appealing to Lemma 3, uninformative events can influence beliefs only if their evocativeness is non-zero. The evocativeness of an uninformative event equals

$$E[x_k a_{kt} | e_t = (0, n_t)] = \frac{1}{2} E[x_k c(0 \Leftrightarrow x_k) | e_t] + \frac{1}{2} E[x_k c(n_k \Leftrightarrow n_t) | e_t]$$

The first term is zero by symmetry of $c(\cdot)$ and the symmetry of the x_k distribution. To evaluate the second term, apply the law of iterated expectations and condition on n_k and n_t :

$$E[x_k c(n_k \Leftrightarrow n_t) | e_t] = E[E[x_k | n_t, n_k] c(n_k \Leftrightarrow n_t) | e_t] = \sigma_{xn} E[n_k c(n_k \Leftrightarrow n_t) | e_t]$$

As Lemma 5 shows, this is non-zero whenever $n_t \neq 0$, and the sign of this term (and hence the event's evocativeness) is $\text{sign}(\sigma_{xn} n_t)$ which establishes the first part. For more general events, the proof is the same: simply note that changes in n_t affect evocativeness whenever $\sigma_{xn} \neq 0$. ■

Even though individuals disregard these events as uninformative, they still influence beliefs by selectively triggering memories. The mediator in this process is σ_{xn} , which determines whether the neutral cue tends to appear with positive or negative information. Take a zero-information but positive n_t event. It recruits other positive n events. The information content of such events depends on σ_{xn} . When σ_{xn} is zero, this tells us nothing about the information content (x) of the triggered memories. Therefore, beliefs will not respond to this noise at all. On the other hand, when σ_{xn} is positive, positive n are associated with positive x , implying that uninformative events with positive n_t trigger memories that convey positive information and lead to larger forecasts. Similarly for negative σ_{xn} .

Let me describe an example that illustrates the effects in Propositions 3 and 4. Consider a political candidate wanting to discredit the incumbent by convincing the electorate that permanent income has dropped. Suppose he begins armed with new, but relatively dull statistics to this end.

How might he transmit this information? He might begin by placing an ad with the statistic alone—“Statistic has gone up...”—and repeat it often. This repetition induces rehearsal which raises memorability. To increase the vividness, he might place the same ad but give the statistic a catchy name (e.g. the misery index). The catchy name provides a cue that makes it more likely to be accessed, thereby increasing vividness. This makes the information contained in the ad more likely to be remembered and influence voters’ beliefs. Alternatively, he might couch the statistic in a heart-wrenching testimonial. An unemployed worker with a strong work ethic paints his difficulties finding work after being laid off from a seemingly secure job. Full of details, this ad will likely trigger the constituency’s own memories of similar stories: “That reminds me of my friend Bob who...” Such an ad alters perceptions of the past and corresponds to having negative evocativeness (it triggers memories which convey negative information about permanent income). Moreover, the next time one hears similar stories, the ad along with the statistic will likely be recalled. Thus, its vividness has also gone up. Finally, he might decide to abandon the statistic completely and simply use the testimony. Even though this would provide little information, it may still influence beliefs by altering which memories are recalled and rehearsed. As before, the ad may remind individuals of other (more informative) stories. In the absence of the statistic, however, future recall of the ad has little effect: it will be discarded as uninformative. This example illustrates how beliefs may respond to vividness, evocativeness and non-information. Experimental work has fleshed out each of these parts as well.

3.5.1 Relation to Experimental Findings: Salience

Many anecdotes suggest the heightened impact of salient information, for which several reasons can be marshaled. The current model simply identifies and formalizes two reasons. I present two canonical experiments that illustrate each of these reasons—vividness and evocativeness.

Thompson, Reyes and Bower (1979) place subjects into the role of jurors, who are asked to read defense and prosecution witness testimony about a drunk driving case. One side’s case was manipulated to be salient while the other’s was manipulated to be pallid. The salience manipulation

was performed through adding inconsequential details to one side’s testimony. For example, in describing the defendant about to leave a party and drive home, the pallid version states that he bumped into a table, and knocked a bowl to the floor. The salient version, on the other hand, states that he knocked a bowl of guacamole dip off a table and onto a white carpet. After reading the two sides, subjects rate the guilt of the defendant and are asked to return the next day. When they return, they are asked to perform the rating again (they do not read the testimony again). Thompson et. al. find that the salience manipulation has *no effect* on the first days’ ratings. The lack of an immediate impact is comforting since it suggests that the salience manipulation did not also manipulate perceived informativeness. For example, we can rule out the possibility that subjects felt that a witness whose testimony contains more details was more reliable. The salience manipulation *did*, however, affect the second days’ judgements of guilt: when the prosecution’s (defense’s) case was more salient, judgements of guilt rose (fell). One interpretation of these results is that the presence of additional cues (guacamole on white carpet) facilitates recall by marshaling associativeness.²⁷ Vividness, as I have defined it, increases because these (irrelevant) cues—for example, spilling something on a carpet—are commonly encountered ones. The increased vividness of one side’s case means that memories over-represent evidence supporting that side.

Hamill, Wilson, and Nisbett (1979) present another experiment, one that resembles evocative-ness more than vividness, by demonstrating a heightened impact of information *contemporaneously*. One set of subjects is presented with a description of a welfare recipient. As Nisbett and Ross (1980, p.57) summarize: “The central figure was an obese, friendly, emotional, and irresponsible Puerto Rican woman who had been on welfare for many years. Middle-aged now, she had lived with a succession of ‘husbands,’ typically also unemployed, and had borne children by each of them. Her home was a nightmare of dirty and dilapidated plastic furniture bought on time at outrageous prices, filthy kitchen appliances, and cockroaches walking about in the daylight. Her children...attended

²⁷A weakness of the current model of the experiment should be pointed out. The guacamole on white carpet cue is effective not because it associates with *current* events but because it associates with past events. In other words, the model needs to allow not only for current events to form associations that facilitate recall, but also memories themselves should form associations that further facilitate recall. I expand on this when I discuss future extensions in Section 5.

school off and on and had begun to run afoul of the law in their early teens, with the older children now thoroughly enmeshed in a life of heroin, numbers-running and welfare.” Another treatment group was given summary statistics on welfare recipients documenting the short median stay (two years) and the small proportion that are on welfare rolls for long periods of time (only 10% for longer than four years). These statistics contrasted sharply with the priors of control subjects.

Both treatment groups and a control group are then asked to state their attitudes about welfare recipients. Treatment subjects receiving the story expressed far more unfavorable attitudes than control subjects, whereas treatment subjects receiving the pallid statistics showed no difference. Evocativeness provide one interpretation of these findings. The story that subjects read is overflowing with cues commonly found in evidence that paints welfare recipients in a poor light—drug use by children, immigrant, obese—whereas the statistics lack such evocative cues. The story thereby triggers evidence from the past that also contain these cues, evidence that will generally be negative. It, therefore, prompts more negative attitudes towards welfare recipients. In this interpretation, it is not that the single case study is taken as informative. Queried, subjects should state that of course they recognize that one story (especially a manufactured one) proves nothing, but that it reminds them of other previously encountered evidence. Of course, the pallid statistics do not possess such cues and, therefore, have lower evocativeness.²⁸ Further, the description of the welfare mother can hardly be considered informative. It, thereby, also illustrates how non-informative signals can have an impact. One implication of this interpretation is that the effect of the manipulation (seeing the story) should disappear over time. If subjects were brought in at later dates, the difference between treatment and control should diminish and eventually vanish.

Contrasting these experiments is a useful way of seeing the difference between vividness and evocativeness. In the guacamole experiment, we added neutral cues that aided recall and thereby saw an impact on beliefs only in the future. This is vividness. In the welfare mom experiment, we added cues that were typically found in negative information and thereby saw a contemporaneous

²⁸That they have *no* effect, however, indicates either that individuals do not put much faith in the statistics (numbers can be manipulated) or that other factors are at play there.

negative impact. This is evocativeness.

3.6 Evocativeness and Over-reaction

The previous propositions illustrate how information content alone does not determine an event's impact; the memories it triggers also matters. Associativeness implies that events trigger memories that convey similar information. Such an effect causes an over-reaction to news: today's events causes similar evidence to be over-represented in memory. The following proposition formalizes this idea.

Proposition 5 *Forecast errors are negatively correlated with the information in the latest event:*

$$\text{Cov}(y_t \Leftrightarrow \hat{y}_t^R, x_t) = \text{Cov}(err_t^R, x_t) < 0$$

The extent of this over-reaction increases with χ and λ :

$$\frac{\partial \text{Cov}(err_t^R, x_t)}{\partial \chi} < 0 \tag{11}$$

$$\frac{\partial \text{Cov}(err_t^R, x_t)}{\partial \lambda} < 0 \tag{12}$$

Proof: Note that $err_t^R = err_t + err_t^m$ and that err_t is independent of x_t . Therefore, $\text{Cov}(err_t^R, x_t) = \text{Cov}(err_t^m, x_t)$. Calculating this:

$$E[err_t^m x_t] = \sum_{k=1}^{t-1} \lambda^{t-k} E[f_{kt} x_k x_t]$$

Using the fact that x_k and x_t are independent, we can write the summand as: $\Leftrightarrow \chi E[x_k x_t a_{kt}]$.

Intuitively, $E[x_k x_t a_{kt}]$ is positive because a_{kt} measures similarity. See Lemma 6. This implies that the overall covariance is negative. To get the comparative statics, let's complete the calculation:

$$E[err_t^m x_t] = \Leftrightarrow \chi E[x_k x_t a_{kt}] (\lambda + \lambda^2 + \dots + \lambda^{t-1}) = \Leftrightarrow \chi E[x_k x_t a_{kt}] \frac{\lambda(1 \Leftrightarrow \lambda^{t-1})}{1 \Leftrightarrow \lambda}$$

Partial differentiation shows that this decreases with χ and λ . The effect of λ is interesting. It happens because when λ is large, the selective sampling of past memories becomes more important, since these memories enter with greater weight into the forecast rule. ■

For example, good information leads to a rosier view of the past, which leads to forecasts that are too large, which leads to a negative forecast error. Over-reaction increases as χ rises because χ quantifies the importance of associativeness. Finally, the effect of λ arises because it measures the importance of history and thereby the importance of selective recall. See Section 3.9.

I have illustrated over-reaction in the politician example described before. The testimony by itself should have no impact but it does. Similarly, the welfare mother experiment suggests over-reaction. Other evidence can be found in studies of the representativeness heuristic by Kahneman and Tversky (1972,1973), Tversky and Kahneman (1971) and Grether (1980). These studies find that in forming assessments individuals place too little weight on base rate evidence and too much weight on the latest piece of information. A similar phenomenon arises in the form of perceptions of a “hot hand”: individuals seeing a streak expect it to continue. While this model does not provide compelling evidence of all the actual experimental evidence (in many of these, the relevant information is directly available and memory plays no role), it generates behavior in real settings that resembles the findings.

3.7 Slow Adjustment

The previous section paints a picture of individuals over-reacting to information. Rehearsal, however, generates under-reaction. To see this, consider an individual who faces an uninformative event $e_t = (0, n_t)$ at time t . Suppose that this event evokes positive memories so that $\mathcal{E}(e_t) > 0$. The results in Prop 4 illustrate how beliefs over-react to this non-information: the positive memories it triggers results in forecasts that are too large. Since these memories are rehearsed, they will experience higher recall probabilities in future periods, meaning that forecasts will continue to be too large. As time goes on, they will decay towards the true value as the effect of the rehearsal on recall probabilities diminishes. To an outsider, the belief changes in later periods will seem as if they were *under*-reaction. At both times $t + j$ and $t + j + 1$, she will see a downward adjustment, as the memories decay in each of those periods. The observer, therefore, sees a negative change followed by another predictable negative change, an apparent under-reaction to the first negative

change. Formally, note from Lemma 4, that:

$$E[\hat{y}_{t+j}^R | e_t = (0, n_t)] = \chi \frac{\lambda}{1 \Leftrightarrow \lambda} \mathcal{E}(e_t) (\lambda \rho)^j$$

Notice that if ρ were zero, this term would be zero, emphasizing the role of rehearsal. If we difference this over time, we find:

$$E[\Delta \hat{y}_{t+j+1}^R | e_t = (0, n_t)] = \chi \frac{\lambda}{1 \Leftrightarrow \lambda} \mathcal{E}(e_t) (\lambda \rho)^j (\lambda \rho \Leftrightarrow 1)$$

which illustrates the negative “drift” in beliefs that follows an over-reaction. In other words, all future belief revisions are negatively proportional to the initial evocativeness. Beliefs will, therefore, appear to drift towards some equilibrium. The intuition behind this finding is that there is more information in her forecast errors than the individual realizes:

$$err_t^R = err_t + err_t^m$$

As with the perfect recall individual, the forecast error tells the forgetful individual that some change has occurred in the permanent component (err_t). But, it also tells the individual the way in which their memory is systematically biased (err_t^m). If she is positively surprised, the forgetful individual should both infer that there probably has been a positive shock and that she is systematically under-sampling positive memories. I discuss this further in Section 3.9.

Slow adjustment arises even more intuitively in a slightly modified version of the model. Suppose that before observing the true event e_t , there is a period where the individual observes a noisy event e'_t (perhaps a rumor). Abstracting away from n_t for now, suppose that x'_t equals x_t plus noise. An example might be the announcement of a government statistic followed by a revision. In this setup, once x_t is revealed the individual should pay no attention to x'_t . But rehearsal combined with associativeness will imply that beliefs will still depend partly on x'_t even after x_t is revealed. Why? Because, even though the individual discards the information contained in x'_t , the set of memories it evoked have been rehearsed and they continue to have an impact in later periods.

Finally, the following proposition shows that forecast errors can be positively auto-correlated.

Proposition 6 *Let $T > t$. When events are very memorable (\underline{f} low, χ and ρ large), then*

$$\text{Cov}[err_t^R, err_{t+1}^R] > 0$$

Proof: (Sketch) I will present a proof for the case where $t \rightarrow \infty$ to abstract from details. The proof for the finite t case is exactly the same but with more constants (that tend to zero as t gets large) involved. The general strategy of the proof is as follows: (1) Use the fact that err_{t+1}^m can be written as a function of err_t^m plus some terms; (2) Substitute into $E[err_{t+1}^m err_t^m]$ to get a $E[err_t^m err_t^m]$ plus some terms that resemble $E[x_k err_t^m]$; (3) these generate opposing signs so that the variance term tends to dominate whenever the probability of forgetting is small.

For the first step in the proof, see Lemma 7 which shows that:

$$err_{t+1}^m = \rho \lambda err_t^m + \sum_{k=1}^{t-1} x_k (\underline{f} \Leftrightarrow \chi a_{k(t+1)})$$

Substitution into $E[err_{t+1}^m err_t^m]$ gives (step 2):

$$\rho \lambda \text{Var}(err_t^m) + \sum_{k=1}^t \lambda^{t+1-k} E[x_k (\underline{f} \Leftrightarrow \chi a_{k(t+1)}) err_t^m]$$

Substitution for the latter gives:

$$\rho \lambda \text{Var}(err_t^m) + \lambda E[(\underline{f} \Leftrightarrow \chi a_{t(t+1)}) x_t err_t^m] + \sum_{k=1}^{t-1} \sum_{j=1}^{t-1} \lambda^{2t+1-k-j} E[x_k x_j (\underline{f} \Leftrightarrow \chi a_{k(t+1)}) f_{jt}]$$

The third term can be written as:

$$\sum_{k=1}^{t-1} \lambda^{2t+1-2k} E[x_k^2 (\underline{f} \Leftrightarrow \chi a_{k(t+1)}) f_{kt}] + \sum_{k=1}^{t-1} \sum_{j=1}^{k-1} \lambda^{2t+1-k-j} \rho^{t-k} E[x_k x_j (\underline{f} \Leftrightarrow \chi a_{k(t+1)}) (\underline{f} \Leftrightarrow \chi a_{jk})]$$

where since x_k and x_j are independent this can be written as:

$$\sum_{k=1}^{t-1} \lambda^{2t+1-2k} E[x_k^2 (\underline{f} \Leftrightarrow \chi a_{k(t+1)}) f_{kt}] \Leftrightarrow \sum_{k=1}^{t-1} \sum_{j=1}^{k-1} \lambda^{2t+1-k-j} \rho^{t-k} E[x_k x_j (\underline{f} \Leftrightarrow \chi a_{k(t+1)}) \chi a_{jk}]$$

Putting these terms together gives:

$$\begin{aligned} \rho \lambda \text{Var}(err_t^m) &+ \sum_{k=1}^{t-1} \lambda^{2t+1-2k} E[x_k^2 (\underline{f} \Leftrightarrow \chi a_{k(t+1)}) f_{kt}] \\ &+ \lambda E[(\underline{f} \Leftrightarrow \chi a_{t(t+1)}) x_t err_t^m] \\ &\Leftrightarrow \sum_{k=1}^{t-1} \sum_{j=1}^{k-1} \lambda^{2t+1-k-j} \rho^{t-k} E[x_k x_j (\underline{f} \Leftrightarrow \chi a_{k(t+1)}) \chi a_{jk}] \end{aligned}$$

Now the first and second terms are clearly positive, where as the third and fourth term are clearly negative. The key insight is that the negative terms tend to zero as memorability gets large (as $\underline{f} \rightarrow 0$) since these predicate on having forgotten x_t or x_k . Therefore, when memorability is sufficiently high, the overall expression is positive. ■

Positive covariance can be understood as overlapping samples. Forgetting is analogous to sampling events from history. Since the samples at times t and T draw from overlapping histories, correlations arise. Moreover, rehearsal implies that memories that were forgotten will be forgotten again, increasing the autocorrelation in forecast errors. The condition that \underline{f} must be sufficiently low occurs for the following reason. Suppose that \underline{f} is very large. Then, the x_t from the past will likely be forgotten and hence x_t shows up with large weight in err_{t+1}^R . We know from Proposition 5 that x_t is negatively correlated to err_t^R . This implies a negative auto-correlation. The next section summarizes on the difficulty of correcting memory errors.

3.7.1 Relation to Experimental Findings: Belief Perseverance and The Curse of Knowledge

The idea that beliefs persevere even after the original evidence has been discredited, or that beliefs adjust slowly also receives much experimental support. Ross, Lepper and Hubbard (1975) provide an interesting experiment along these lines. Subjects are asked to assess the authenticity of suicide notes, where some are real and others manufactured.²⁹ They read these notes one by one, make an assessment of authenticity and then after each one are told whether their assessment is correct. What they do not know, however, is that the feedback is random, completely unrelated to their actual performance. At the end of the task, subjects are asked to evaluate their ability on this task. The first finding is that subjects do indeed respond to the performance information: those who were told that they performed well also perceived themselves to be more able. After this, subjects are told about the complete randomness of the feedback. They are then given a post-experiment questionnaire, where one question on it asks them to reassess their ability on the task.

²⁹Many other experiments have been performed of this type, some using more standard judgement tasks such as logic puzzles.

Ross and his colleagues find that *even after the debriefing*, even after being told that the feedback was completely random, the performance information continues to have impact. Subjects in high (low) performance conditions continue to believe that they are more (less) capable. The current model interprets these findings as the result of rehearsal. As subjects receive positive feedback they come to remember all the qualities that make them good at the task. ‘I am a very empathetic person’, they might think. Events like ‘Just the other day, I helped my best friend through a tough problem’ might occur to them. These memories have been freshly rehearsed and, therefore, even after the initial evidence has been discredited, it continues to have an impact.³⁰

A similar phenomena is found by Camerer, Loewenstein and Weber (1990) who find that subjects are unable to discount superior information when forming forecasts of *others’* forecasts. In their experiment, one group of subjects is asked to predict a firm’s dividends. Another, the treatment, is asked to predict this group’s predictions. The treatment group is divided into two, with one having superior information—they know the true value of dividends—while the other group has no extra information. They find that the group with the same information as the individuals they are predicting is fairly accurate in forecasting the forecasts. The group with the superior information, however, does poorly. They expect forecasts that are systematically too close to the true value, which they know. A memory interpretation is that even though individuals discount the true value, their memories still oversample events supportive of the true value. In other words, the associations formed by the information are not undone.³¹ Notice the similarity to belief perseverance. Both point at the inability of the individual to discard information.

3.8 Opposing forces in Belief Dynamics

The previous sections illustrate two conflicting experimental findings, both of which show up in the model. On the one hand, individuals respond too much to new information, and on the other, their

³⁰Nisbett and Ross (1980) also informally suggest this interpretation. An alternative reading is that individuals do not believe the debriefing. Verbal reports would allow one to easily separate these two interpretations.

³¹When individuals are asked to backcast what their forecasts were or would have been, a similar phenomenon arises. Once the outcome is known, individuals over-estimate their propensity to have forecasted it. The interpretation in this model is the same as with the curse of knowledge. Once the true value is known, it is hard to discount it. See Fischhoff (1982).

forecast errors may be positively autocorrelated. One advantage of a model such as this is that it allows us to trade off such effects and figure out when we expect one to arise over the other. To this end, I examine how belief changes depend on lagged information. This would require that we compute: $Cov[\Delta\hat{y}_T^R, \Delta y_t] = Cov[\hat{y}_T^R \Leftrightarrow \hat{y}_t^R, \Delta y_t]$ A positive coefficient implies an “under-reaction” of sorts: a positive value in the past means that individuals are revising their beliefs upwards today. On the other hand, a negative coefficient implies an “over-reaction” of sorts: a positive piece of information in the past means that individuals are revising their beliefs downwards today.

Proposition 7 *Suppose that forgetting probabilities are small, so that \underline{f} is high and ρ and χ are low. Then:*

$$Cov[\Delta\hat{y}_{t+1}^R, \Delta y_{t-1}] < 0 \quad (13)$$

When these probabilities are large, however:

$$Cov[\Delta\hat{y}_{t+1}^R, \Delta y_{t-1}] > 0 \quad (14)$$

When the covariance is negative, then a change in λ makes it more negative:

$$\frac{\partial Cov[\Delta\hat{y}_{t+1}^R, \Delta y_{t-1}]}{\partial \lambda} < 0 \quad (15)$$

Proof: Now,

$$\Delta\hat{y}_{t+1}^R = \Delta\hat{y}_t \Leftrightarrow err_{t+1}^m + err_t^m$$

and \hat{y}_t is independent of all lagged information. Therefore, the covariance equals:

$$E[err_t^m \Delta y_{t-1}] \Leftrightarrow E[err_{t+1}^m \Delta y_{t-1}]$$

From Lemma 7, we can write err_{t+1}^m in terms of err_t^m . Substituting for this gives:

$$(1 \Leftrightarrow \lambda \rho) E[x_{t-1} err_t^m] \Leftrightarrow \sum_{k=1}^t \lambda^{t-k+1} E[(\underline{f} \Leftrightarrow \chi a_{k(t+1)}) x_k x_{t-1}]$$

Reapplying Lemma 7 to err_t^m gives:

$$(1 \Leftrightarrow \lambda \rho) \lambda \rho E[x_{t-1} err_{t-1}^m] + (1 \Leftrightarrow \lambda \rho) \sum_{k=1}^{t-1} \lambda^{t-k} E[x_k x_{t-1} (\underline{f} \Leftrightarrow \chi a_{kt})] \Leftrightarrow \sum_{k=1}^t \lambda^{t-k+1} E[(\underline{f} \Leftrightarrow \chi a_{k(t+1)}) x_k x_{t-1}]$$

Note that x_k and x_{t-1} are independent in the summations for $k \neq t \Leftrightarrow 1$, leaving:

$$(1 \Leftrightarrow \lambda \rho) \lambda \rho E[x_{t-1} err_{t-1}^m] + (1 \Leftrightarrow \lambda \rho) \lambda E[x_{t-1}^2 (\underline{f} \Leftrightarrow \chi a_{(t-1)t})] \Leftrightarrow \lambda^2 E[x_{t-1}^2 (\underline{f} \Leftrightarrow \chi a_{(t-1)t})]$$

Define C to be $E[x_{t-1}^2(\underline{f} \Leftrightarrow \chi a_{(t-1)t})]$ which also equals $E[x_{t-1}^2(\underline{f} \Leftrightarrow \chi a_{(t-1)(t+1)})]$. This gives:

$$(1 \Leftrightarrow \lambda \rho) \lambda \rho E[x_{t-1} \text{err}_{t-1}^m] + C \lambda (1 \Leftrightarrow \lambda (1 + \rho))$$

Substituting for the first part from the proof of Proposition 5 gives:

$$\Leftrightarrow \chi \frac{\lambda (1 \Leftrightarrow \lambda^{t-1})}{1 \Leftrightarrow \lambda} E[a_{kt} x_k x_t] + C \lambda (1 \Leftrightarrow \lambda (1 + \rho)) \quad (16)$$

Suppose events are very memorable, so that the forgetting probability, \underline{f} is low and χ and ρ are high. Then $(\lambda (1 \Leftrightarrow \lambda (1 + \rho))) C = (\lambda (1 \Leftrightarrow \lambda (1 + \rho))) E[x_{t-1}^2(\underline{f} \Leftrightarrow \chi a_{(t-1)t})]$ is small or even negative. The first term is already negative, so that in this case the correlation is negative. Suppose, on the other hand, that events are easy to forget so that \underline{f} is high and χ and ρ are low. Then, the first term tends to zero, while the second term implying a positive correlation.

Differentiating with respect to λ gives:

$$\Leftrightarrow \chi \frac{1 \Leftrightarrow \lambda^{t-1}}{1 \Leftrightarrow \lambda} E[a_{kt} x_k x_t] + C (1 \Leftrightarrow 2\lambda (1 + \rho))$$

which is the same as equation (16) except (i) it has been divided through by λ and (ii) $1 \Leftrightarrow \lambda (1 + \rho)$ has been replaced by $1 \Leftrightarrow 2\lambda (1 + \rho)$. The first has no result on the sign and the second only makes it more negative since C is positive and $1 \Leftrightarrow \lambda (1 + \rho) > 1 \Leftrightarrow 2\lambda (1 + \rho)$. Consequently if equation (16) is negative the derivative with respect to λ is also negative.

■

The intuition behind this proposition is that there are two effects that govern belief dynamics: forgetting and over-reaction. On the one hand, yesterday's information may have been forgotten, which means that the individual must learn it again. This induces a positive correlation between beliefs and lagged information. On the other, over-reaction means that the individual responded too much to x_t when it occurred, meaning that beliefs must correct for this. This induces a negative correlation. When events are on average quite memorable, the over-reaction effect dominates.

When events are readily forgotten, the individual must relearn old information. The dependency on λ reflects the discussion in Section 3.9. A greater emphasis on history implies over-reaction is larger and takes a longer time to undo.

3.9 The Role of λ

Recall that λ measures the weight put on past values in the forecast. In this section I will examine how λ mediates over-reaction and slow learning. I will argue that λ can be measured easily and therefore the empirical tests involving λ can actually be implemented.

Let's begin by considering the impact of forgetting an event. The memory error equals:

$$err_t^m = \hat{y}_t \Leftrightarrow \hat{y}_t^R = \sum_{k=1}^{t-1} \lambda^{t-k} (1 \Leftrightarrow R_{kt}) x_k = \sum_{k=1}^{t-1} \lambda^{t-k} (F_{kt}) x_k$$

If $F_{kt} = 1$, so that event e_k is forgotten, the memory error would go up by $\lambda^{t-k} x_k$. This shows that the impact of forgetting an event declines as time passes ($t \Leftrightarrow k$ gets large). Moreover, the rate of decline depends on λ . The larger is λ , the larger is the effect of forgetting an event in the distant past. Why does this happen? As time proceeds, the information lost due to forgetting x_k is slowly re-learned through the y_t . Events provide perfect signals of the permanent shocks, so forgetting them means that this perfect signal is lost. In the absence of this signal, the individual still learns about the permanent shock but this time through y_t . Since y_t is noisy, however, this learning is slow. The more distant the memory, the more time there has been to learn about the event through observations of y_t instead. This establishes why there will be slow learning. This learning occurs at rate λ because λ measures the noise-signal ratio in y_t . When it is large, y is a very noisy signal of permanent income and forgotten events are learned about very slowly. To summarize, λ captures how quickly a forgotten event can be relearned through the y , and hence how quickly errors in memory are corrected.

Let's now return to rederiving how beliefs respond to an event e_t :

$$E[\hat{y}_t^R | e_t] = x_t \Leftrightarrow \sum_{k=1}^{t-1} \lambda^{t-k} E[x_k f_{kt} | e_t]$$

$$x_t + \sum_{k=1}^{t-1} \lambda^{t-k} (\chi E[x_k a_{kt} | e_t])$$

$$x_t + \chi \mathcal{E}(e_t) (\lambda + \lambda^2 + \lambda^3 + \dots + \lambda^{t-1})$$

To get the second equation, we exploit the fact that f_{x_k} is independent of x_t as is $f_{k(t-1)} x_k$. The third equation comes from the definition of $\mathcal{E}(e_t)$. To interpret this equation, notice that at time k , associativeness results in a selective sampling that has effect equal to the evocativeness, $\mathcal{E}(e_t)$. But as we've seen the impact of recall mistakes on beliefs depends on λ . In the formula, we see that selectively recalling the events at time k has impact $\lambda^{t-k} \mathcal{E}(e_t)$. Taking $t \rightarrow \infty$ for simplicity, the impact of selective recall is:

$$\chi \mathcal{E}(e_t) (\lambda + \lambda^2 + \lambda^3 + \dots) = \chi \mathcal{E}(e_t) \frac{\lambda}{1 \Leftrightarrow \lambda}$$

Therefore, as λ increases, the importance of selective recall increases. Intuitively, when λ is large, the triggering of certain types of memories over others has bigger impacts because the past matters more.

We, therefore, see two basic properties of λ . It both measures extent of over-reaction and how slowly individuals adjust their memory mistakes. These two observations are especially interesting since λ can be measured in standard data sets. Carroll and Samwick (1995) present a technique that basically allows us to estimate λ . First, notice that

$$\Delta^d y_t = y_t \Leftrightarrow y_{t-d} = \nu_t + \nu_{t-1} + \dots + \nu_{t-d} + \epsilon_t \Leftrightarrow \epsilon_{t-d}$$

So that:

$$Var(\Delta^d y_t) = d\sigma_\nu^2 + 2\sigma_\epsilon^2$$

Therefore, one can compute $\Delta_d y_t$ in the data for many different d . A regression on d then allows one to back both variances out. Formally let s_{id} be the variance for firm or individual i at horizon d . Let's focus on all firms in an industry or individuals in an industry or occupation and assume that λ is common within this cell. Then we can form s_{id} and run the OLS regression:

$$s_{id} = \alpha + \beta * d$$

Then $\frac{\alpha}{2} = \sigma_\epsilon^2$ and $\beta = \sigma_\nu^2$ which allows us to compute λ . Since we can run this regression cell by cell we can get a different λ for each group (industry or occupation).

Of course, in practice, several difficulties may arise. Suppose y_t represents individual earnings. Effective estimation of λ may require panel data of longer time periods or higher quality than currently available, though Carroll and Samwick (1995) perform a related exercise on the PSID with some success. Even were this data available, several other problems remain. The true stochastic process for y_t may not match the $AR(1)$ plus white noise process assumed. For example, normality of the error terms may fail due to outliers (measurement error), or fat tails. Such problems may be less severe in other applications—such as with firm earnings—where measurement error is less of a problem, though of course they still remain. One should keep in mind that tests of the proposition, while feasible, will require some care.

3.10 Ignoring One’s Memory

We have so far discussed the many biases that can arise from the faultiness of memory. This raises the natural question, what if the individual were simply to ignore his memories and rely solely on the hard data? He might say, I know my memory is faulty, so I pay no attention to it. Moreover, an answer to this question may give us some guidance on when we would want to use expert systems, more mechanical decision rules, to replace or augment boundedly rational decision makers.

To answer this question, consider a decision rule, \hat{y}_t^I , one that simply ignores all memories:

$$\hat{y}_t^I = x_t + \sum_{k=1}^{t-1} (1 \Leftrightarrow \lambda^{t-k}) \Delta y_k$$

This rule is equivalent to a naive individual who forgets everything (all $R_{kt} = 0$).³² The forecast

³²This rule naturally extends the naive forecast to the memory ignorance case. It should be noted, however, that this rule is not the optimal one when ignoring memory. Optimality would require that individuals recognize that there is more signal in the Δy_t now that events are not known. An intuitive way of seeing this is that the effective λ should be lower. Further complications, however, make analysis of the optimal rule a great deal more complicated. The allowance of forgotten events means that the information being learned is no longer normal (events have a discrete component—they occur with probability p). Non-normality in turn complicates the forecasting rule, making it non-linear in Δy_t . The simple rule studied in this section can be thought of as accompanying the naive case, the focus of this paper. The more complex one may also be worthy of study and would most correctly accompany study of the sophisticated case.

error in this case can again be written as:

$$err_t^I = err_t + \sum_{k=1}^{t-1} x_k$$

Define $err_t^i = \sum_{k=1}^{t-1} x_k$ as the analogue of err_t^m . Thus, while the naive individual experiences a distortion in their forecast error of err_t^m , the one who ignores memory experiences a distortion equal to err_t^i .

Let's now measure the relative performance of these rules (naive forecasts that use memory against those that ignore memory all together) on two dimensions: bias and variance of the forecast errors. To assess bias, consider the expected forecast error conditioning on today's event e_t :

$$\begin{aligned} E[err_t^I|e_t] &= 0 \\ E[err_t^R|e_t] &= \Leftrightarrow \chi \mathcal{E}(x_t) \frac{\lambda}{1 \Leftrightarrow \lambda} (1 \Leftrightarrow \lambda^{t-1}) \end{aligned}$$

The naive individual is biased because of associativeness. By discarding his memories, however, the ignorant one has no bias. Associativeness has no chance to operate. Let's now consider the variances:

$$Var[err_t^I|e_t] = Var[err_t|e_t] + Var[\sum_{k=1}^{t-1} \lambda^{t-k} x_k] = Var[err_t] + p\sigma_x^2 \frac{\lambda^2}{1 \Leftrightarrow \lambda^2} (1 \Leftrightarrow \lambda^{2(t-1)})$$

The ignorant individual suffers a large loss in terms of variance because a lot of information is omitted from the forecast. The variance of the naive individual is:

$$Var[err_t^R|e_t] = Var[err_t|e_t] + Var[err_t^m|e_t]$$

By using more information, the naive individual should intuitively have lower variance. While this is true for many parameters, it is possible that using memory results in *greater* variance. Associativeness tends to recall information all of the same kind. Therefore, reliance on memories can result in a very volatile estimator. See Lemma 8. To summarize, ignoring information results in less bias and possibly greater variance.

Suppose that χ is very small. Then the naive individual is close to being unbiased. Moreover, according to Lemma 8 in the appendix, we see that she has lower variance than if she were to

ignore her memories. Therefore, for small χ she is better off using her memories. As χ increases, both the bias and variance increase for the naive individual (see Lemma 8). Therefore, the larger χ , the more likely it is that ignoring information will be optimal.

We can also perform comparative statics on the relative importance of bias and variance. Let's focus on the case of moderate χ , so that the naive forecast is better in terms of variance, but is worse in terms of bias. Therefore, the optimality of using one's memory will increase with the relative costs of large variance versus large bias.

3.11 Summary of Results

The main results of the model are:

1. On average and in the long run, forecasts are unbiased under a symmetry assumption;
2. Events with greater vividness have higher steady state recall probabilities;
3. Both increased vividness and evocativeness increase an event's impact on beliefs;
4. Neutral components can affect beliefs;
5. Individuals over-respond to news;
6. Beliefs adjust slowly;
7. λ measures (roughly) the proportion of transitory to permanent shocks. As λ increases (i) over-reaction increases and (ii) beliefs adjust more slowly;
8. When events are (on average) very memorable, changes in beliefs are negatively correlated with lagged information. When they are not, changes in beliefs are positively correlated with lagged information.
9. Even with a faulty memory, it makes sense to use it as long as χ is small. As the cost of being biased increases, however, one may do better by ignoring memories and focusing on hard data.

4 Applications

4.1 The Consumption Decision

Let i index individuals, and y_{it} represent and income, c_{it} denote consumption and $u(c)$ be the instantaneous utility function taken to be the same across individuals. Suppose the individual maximizes discounted (subjective) expected utility, where the discount rate is δ . Assume that she faces no borrowing or savings constraints and can borrow or save risklessly at a rate r , and that $\delta = \frac{1}{1+r}$. Further, impose a no-Ponzi game condition so that there is no infinite borrowing. Under these conditions, marginal utility of consumption will be equated: $u'(c_t) = u'(c_T)$ for all t, T . Taking a quadratic or log-utility function implies that consumption will be equalized across time. We can write time t consumption, therefore, as a function of time t assets and the expected income in future periods. In the current model, an individual's forecast of time $t+k$ income at time t is the same for all k . This allows us to write consumption as:

$$c_{it} = \frac{r}{1+r} A_{it} + \hat{y}_{it}^R$$

where $A_0 = 0$ and $A_{i(t+1)} = (1+r)(A_{it} + y_{it} \Leftrightarrow c_{it})$ is the assets. Differencing across time gives:

$$\Delta c_{it} = \frac{1}{1+r} \Delta \hat{y}_{it}^R + y_{i(t-1)} \Leftrightarrow \hat{y}_{i(t-1)}^R$$

In other words, the change in consumption is proportional to the change in income expectations plus the time $t \Leftrightarrow 1$ forecast error. This is intuitive since permanent income considerations completely determine consumption in this model. Substituting for the forecast error gives:

$$\Delta c_{it} = \frac{1}{1+r} \Delta \hat{y}_{it}^R + err_{i(t-1)}^R$$

Now, suppose that the income process is the sum of two components: one specific to the individual and an aggregate component. Letting \bar{y}_t be the aggregate component, and y_{it}^0 be the individual specific one, we write:

$$y_{it} = y_{it}^0 + \alpha_i \bar{y}_t$$

where α_i measures how much the aggregate shock influences the individual. Both the aggregate and individual income components follow processes described so far and the individual income components are iid across people.³³ Events are observed for both processes. Let \bar{c}_t be aggregate consumption.

In this simple Permanent Income setup, consumption changes should be unpredictable. Since they essentially represent belief changes, one should not be able to predict them on the basis of lagged information available to consumers. In contrast, the errors of the forgetful forecaster lead to consumption predictability, and the pattern of this predictability can be pinned down under certain conditions.

Prediction 1 *Suppose:*

1. *Personal events are highly memorable and aggregate events are not very memorable ; and*
2. *α_i is small*

then at the micro level:

$$\begin{aligned} \text{Cov}(\Delta c_{i(t+k)}, \Delta y_{it}) &< 0 \\ \frac{\partial \text{Cov}(\Delta c_{i(t+k)}, \Delta y_{it})}{\partial \lambda_i} &< 0 \\ \frac{\partial \text{Cov}(\Delta c_{i(t+k)}, \Delta y_{it})}{\partial \alpha_i} &> 0 \end{aligned}$$

while at the aggregate level:

$$\text{Cov}(\Delta \bar{c}_{t+k}, \Delta \bar{y}_t) > 0$$

To see how this prediction works, note that:

$$\text{Cov}(\Delta c_{i(t+k)}, \Delta y_{it}) = E[\Delta \hat{y}_{i(t+k)}^R \Delta y_{it}] + E[err_{i(t+k-1)}^m \Delta y_{it}]$$

Taking the first term, we can break it into the components due to the aggregate shock and the parts due to the idiosyncratic component:

$$E[\Delta \hat{y}_{i(t+k)}^R \Delta y_{it}] = E[\Delta \hat{y}_{i(t+k)}^{0R} \Delta y_{it}^0] + \alpha_i^2 E[\Delta \hat{y}_{i(t+k)}^R \Delta y_{it}^0]$$

³³There is a slight oddness in the results here. Income is normally distributed meaning that it might well be negative. Using a log-normal distribution would generate all the results here but with added technical complications. The goal here is simply to illustrate the kinds of results that arise rather than to flesh out a structural model.

where because of independence, I have dropped terms such as $E[\Delta\hat{y}_{i(t+k)}\Delta\bar{y}_t]$. Applying Proposition 7, we know that the first term here is negative (we have assumed that personal events are very memorable), and that the second term is positive (we have assume that aggregate events are easily forgotten). Therefore, if α_i is small, the whole expression is negative. The second term in the expression is:

$$E[err_{i(t+k-1)}^{0m}\Delta y_{it}^0] + \alpha_i^2 E[e\bar{r}r_{i(t+k-1)}^m\Delta\bar{y}_{it}]$$

where err_{it}^{0m} is the memory error for the idiosyncratic income component and $e\bar{r}r_{it}^m$ is the memory error for the aggregate component. Just as in the proof of Proposition 7, these correlations are negative when events are memorable and positive when events are easy to forget. Therefore, the first term here is negative and the second term is positive with the smallness of α_i generating a negative sign for the sum. Putting this all together gives that, under the assumptions we have made,

$$Cov(\Delta c_{i(t+k)}, \Delta y_{it}) < 0$$

The partial with respect to λ_i come clearly from Proposition 7, whereas the partial with respect to α_i comes from the fact that the aggregate contribution to the covariance is positive.

Suppose now that we aggregate up consumption and income. Since the idiosyncratic components of income and its forecasts are iid across people, aggregation produces zero for these. This gives:

$$Cov(\Delta\bar{c}_{(t+k)}, \Delta\bar{y}_t) = \bar{\alpha}^2 E[\Delta\hat{y}_{(t+k)}^R\Delta\bar{y}_t] + \bar{\alpha}^2 E[e\bar{r}r_{i(t+k-1)}^m\Delta\bar{y}_t]$$

where $\bar{\alpha}$ is the average of α_i . Reapplying Proposition 7 as before tells us that this term will be positive. This establishes the aggregate results.

Intuitively, over-reaction dominates for the idiosyncratic components of income since these are memorable. The dominant effect is that individuals over-react to their private information. Their boss calls them in, tells them that they have a bright future, and this causes them to selectively recall other information that makes them think they have high ability, and hence, high permanent income. At the micro level, the smallness of α_i guarantees that the reaction to the aggregate information does not matter. As one aggregates up, the idiosyncratic over-reactions cancel out.

Macro-covariances, therefore, depend on recall of the aggregate component. Because aggregate information is forgotten, there is under-reaction to it. This leads to a positive covariance at the aggregate level.

The first assumption of differential memorability can be justified only by appeal to intuition (or perhaps through surveys): personal events may hold more memorability for consumers because they deal with many more everyday events than aggregate events. The second assumption receives some support in the data, as Pischke (1995) and others have argued that the aggregate component of individual income is small.

At the micro level, the first part of this prediction resembles “rule of thumb” consumers, ones who consume more of their income than permanent income considerations would justify. The prediction has generally found support in the literature. In their seminal study, Hall and Mishkin (1982) found compelling evidence in the PSID for a negative correlation between consumption changes and lagged income changes. Later work by Mariger and Shaw (1990), also with the PSID, but including later years, fails to find it. Of course, given the noisiness of this data, statistical significance always remains a problem. Results from other data sets and other countries do generate results consistent with a negative correlation, such as Hayashi (1985a,1985b) and Jappelli and Pagano (1988). An alternative interpretation of this finding, of course, is liquidity constraints. The second and third predictions, however, have not been tested as far as I know. Finally, the macro prediction has received support, as seen in Campbell and Mankiw (1989). Deaton (1992) summarizes this evidence.

Now, suppose that we go back to a single individual, set $\alpha_i = 0$, and allow for several income streams. The marginal propensity to consume out of these different income streams will depend on the extent of the that stream’s evocativeness. Note, from Proposition 3, that the stronger the recruitment effect the larger the forecast error and hence stronger the mean reversion. Define y_{st} to be income stream s and MPC_s to be marginal propensity to consume out of stream s . Then:

Prediction 2 *In general $MPC_s \neq MPC_{s'}$. Moreover,*

$$MPC_s > MPC_{s'} \Rightarrow Cov(\Delta c_t, \Delta y_{s(t-1)}) < Cov(\Delta c_t, \Delta y_{s'(t-1)})$$

To see, how this works note that:

$$MPC_s = Cov(\Delta c_t, \Delta y_{st}) = E[\Delta \hat{y}_{st}^R \Delta y_{st}] + E[err_{s(t-1)}^m \Delta y_{st}]$$

Since $err_{s(t-1)}^m$ is independent of Δy_{st} we can drop the second term. This leaves us with the first term, which we can write as:

$$E[\Delta \hat{y}_{st} \Delta y_{st}] \Leftrightarrow E[\Delta err_{st}^m \Delta y_{st}]$$

The first term here is the appropriate MPC in the absence of any memory mistakes. The second term represents the distortion:

$$E[err_{st}^m \Delta y_{st}] = \chi E[\mathcal{E}(e)x] \frac{\lambda}{1 \Leftrightarrow \lambda}$$

This will in general be different for different income streams especially since $E[\mathcal{E}(e)x]$ will vary. In other words, streams that have high evocativeness, where information about earnings in that stream relies heavily on soft information that has many cues, will have larger MPCs. The implication for greater negative lagged correlation comes directly from the discussion to date. The greater $E[\mathcal{E}(e)x]$, the greater the over-reaction and hence the greater the correlation to lagged income changes.

Intuitively, the prediction follows because changes in different income streams invoke different “visceral” reactions. Empirically, differences in MPC has received some support (Thaler, 1990). Serious empirical difficulties arise, however. Empirical differences in MPCs may represent true differences in propensities to consume permanent income. Alternatively, they may represent differences in the informativeness of income changes. Yet another possibility is that they may represent differences in information between the econometrician and the individual due either to measurement error or private information. This makes testing such predictions heavily reliant on structural assumptions about the income process. On the other hand, the relationship between MPC and excess sensitivity has not been tested as far as I know, and the empirical difficulties here may be less severe.

4.2 Asset Prices

Take d_t to be a firm's earnings, which are paid out as dividends at time t . Assume that there are an infinity of such firms whose earnings are independent so that all risk can be diversified away. Furthermore, take the stochastic process of d_t to be the same as y_t . Again, assuming away bubbles, prices equal the net present value of expected dividends:

$$p_t = \frac{\hat{d}_t^R}{r}$$

The first prediction states that prices are correct on average:

Prediction 3 *On average and in the long run, prices equal the net present value of dividends.*³⁴

$$\lim_{k \rightarrow \infty} E[p_{t+k} | h_t] = \frac{E[d_{t+k} | h_t]}{r} \quad (17)$$

This follows intuitively from the unbiasedness of forecast errors in Proposition 2.

Even though prices are on average correct, they may move around too much. As we have seen in Propositions 3 and 4, beliefs may react to non-information or over-react to new information.

This generates:

Prediction 4 *Stock prices move "too much", i.e. for no apparent reason. When χ is large enough, $Var(p_t) > Var(\frac{d_t}{r})$. The larger λ the larger this excess volatility.*

The intuition here is simple. Volatility is large for two reasons. First, evocativeness implies that prices respond too much to new information. For example, a story about a firm having to recall a product may evoke other memories on the part of traders, such as the time the management had to cancel an expensive investment, or other run-ins with regulators. These evoked memories lead to a larger reaction to this news than is merited by the information it conveys. Second, and less interestingly, the randomness of recall itself produces volatility.³⁵ Both of these imply that prices

³⁴In the model, asset prices are determined solely by forgetful agents' expectations. In reality, arbitrage conditions also play a role, as agents who possess better memories or better records will attempt to profit from the mispricings. Nevertheless, limits on arbitrage may allow such mispricings to persist as argued in papers such as De Long, Shleifer, Summers and Waldmann (1990) and Shleifer and Vishny (1997). Of course, whether they survive or not is an empirical question. This paper outlines the nature of mispricings under the joint hypothesis of limited arbitrage and forgetfulness. Such a well specified model alleviates data snooping worries since it attempts to predict price anomalies before examination of the data (Lo and MacKinlay, 1990)

³⁵This becomes non-existent when we allow for many individuals trading, since the sheer randomness of recall will wash out in the aggregate. Of course, recall *biases* will not since the individuals' observe the same events.

will move even on days where there seems no news. Shiller (1989) gives a plethora of evidence of excess volatility. Cutler, Poterba and Summers (1989) examine stock price movements and argue that large price movements are often unaccompanied by any significant news. Roll (1984) performs a similar exercise for orange juice futures and finds that changes in weather forecast account for little of the variance in price movements, even though this is by far the biggest source of information. The relationship to λ comes from the discussion in Section 3.9 that over-reaction increases with λ .

With regards to the actual dynamics of prices, Propositions 5, 6 and 7 tells us that prices may not mean revert. A positive change in prices might either reflect the undoing of a previous memory error or an over-reaction to new information. In the former case we expect price drift, while in the latter we expect mean reversion. Notice, however, that in this model prices minus the dividends (divided by r) equals:

$$p_t \Leftrightarrow \frac{d_t}{r} = \Leftrightarrow \frac{err_t^R}{r}$$

Recall that $err_t^R = err_t + err_t^m$ so that when prices are high relative to dividends, the memory error is likely negative. Consequently, prices must be going up. Finally, we have seen that when λ is large, learning about the memory error is slower.

Prediction 5 *Mean reversion does not necessarily hold. But, there is conditional mean reversion, so that $p_t \Leftrightarrow \frac{d_t}{r}$ forecasts future returns:*

$$Cov(p_{t+k} \Leftrightarrow p_t, p_t \Leftrightarrow \frac{d_t}{r}) < 0$$

for large k . The larger λ the slower the mean reversion.

Campbell and Shiller (1988a,1988b) present evidence on mean reversion in the aggregate data. Lakonishok, Shleifer and Vishny (1994) present evidence at the micro level. Again, the prediction on λ is new as far as I know.

Finally, we can ask, how do prices respond to earnings announcements? Earnings announcements in this model correspond to the revelation of d_t (y_t in the original analysis). Understanding this process will require a more careful analysis. Currently, we only examine belief revisions around events and d_t announcements jointly. This is appropriate in the consumption context, for example, where we use time-aggregated data. With financial market data, however, we can separately

observe belief revisions around d_t announcements. Modeling these revisions requires assumptions about how recall operates when d_t (as opposed to events) is announced. Do these announcements trigger memories of any events? Suppose as a first pass that there is no associativeness effect. As we have seen in Section 3.9, d_t announcements are times when the individual learns about her memory error. From there, we know that this learning is slow so that belief revisions on earnings announcement will be too small. In other words, following earnings announcements, there will be under-reaction and price drift.

Now, let's complicate the story and allow for some association effect so that earnings changes themselves can serve to trigger memories. We then have two opposing effects. From above, if this association effect is sufficiently small, we will get under-reaction. If the effect is sufficiently large however, we will get over-reaction through the same logic as before: positive earnings trigger supporting memories. The model thereby produces an ambiguity. Some headway can be made by considering *sequences* of consistent earnings announcements. Consider observation of one positive earnings innovation. As noted, this may not result in over-reaction if the association effect is small. But with a sequence, each announcement will trigger and rehearse positive past events. As the sequence gets large, the cumulated effect of associativeness gets larger and larger. Eventually, this will lead positive memories to be over-sampled and over-reaction. Consequently, we can compute an n^* such that n^* announcements result in under-reaction, but $n^* + 1$ results in over-reaction. Notice in this prediction that n^* might be 0, in which case we will always have over-reaction. This happens whenever the initial association effect of earnings announcements is sufficiently large. Alternatively, n^* may be infinity if the association effect is sufficiently weak.

Prediction 6 *There exists $n^* \geq 0$ such that: (i) Conditional on n^* or less consistent earnings announcements, prices will exhibit drift or under-reaction; and (ii) Conditional on $n^* + 1$ or more consistent earnings announcements, prices will exhibit mean reversion or over-reaction. The extent of drift and mean-reversion is larger as λ increases.*

A body of empirical work, summarized in Bernard (1993) and Barberis, Shleifer and Vishny (1997) indicates that $n^* > 0$. In general, prices exhibit under-reaction to earnings announcements. Conditioning on several earnings announcements generates over-reaction. Again, the relation to λ has

not been tested.

The relation to λ that permeates all the financial market predictions is worth discussing as we conclude this application. What kind of firms do we *ex ante* expect to have large λ ? Taken literally, these are firms that have a high ratio of transitory to persistent shocks. In practice, however, the $t \rightarrow \infty$ condition will not hold. In the finite case, the information content of d_k (and hence the value of λ) will depend on the priors. For example, large and well-established firms will not use much information from d_k because priors about them are relatively precise. Small or new firms will be the ones for which d_k information plays a larger role. Earnings information will be more important in determining value for these firms. Consequently, the above implications can be translated into saying that mispricings are larger and undone more slowly for small or new firms.

5 Conclusion

To summarize, this paper has built a simple model of memory limitations. The model has been based on two basic facts drawn from scientific research on the topic: rehearsal and association. Interestingly, these two facts in combination generate several of the experimentally found biases in decision making under uncertainty. This suggests that memory limitations might be an important component for realistic models attempting a unified treatment of bounded rationality. The model also generate relevant predictions in the economic applications we have examined: consumption and asset pricing. We have also seen how previously untested predictions arise. Many other applications are possible that have not been pursued here: advertising, subjective performance evaluation (where assessments of an individual may depend on intangible aspects of past performance), and bargaining situations (where opponents may disagree on the past) are a few of the examples. Each of these has its own subtleties.

Let me conclude by outlining three directions of my current work. First, this paper has focused on the naive case. What does behavior in the sophisticated case look like? I have already given a flavor of the kinds of results that might arise in footnote 25. As pointed out there, the deviations from full rationality become no less interesting. Another point to be made here is that in the

case of outsiders manipulating memory limitations even if the mean effect is “taken out” due to sophistication, the possibility for manipulation can still have real effects. For example, if firms attempt to use advertising to manipulate memories but individuals attempt to undo it, the Nash Equilibrium can result in positive levels even though there will be no *equilibrium* distortion in beliefs. In other words, a standard “signal jamming” argument can be applied when advertising attempts to manipulate sophisticated players.

Second, associativeness as formulated in this paper has a failing. While current events can trigger related memories, the memories that one recalls cannot themselves trigger other memories, an extension I refer to as association chains. Allowing for such chains raises the possibility of multiple steady states in recall. Consider a world in which there are only two types of events, good and bad. For a fixed history, one possible steady state is that good events by chance have had high recall and bad events have had low recall. By rehearsal, good events also have high current recall probabilities r_{kt} . Such an individual appears optimistic since he systematically over-recalls good events. Moreover, when he encounters a good event, it will have higher recall in the future. The existing stock of good events have high recall and will, therefore, trigger this *new* event frequently through association chains, generating a great deal rehearsal and raising its steady state recall probability. Similarly, a bad event, by virtue of its association chains being with low recall probability bad events will tend towards a low recall steady state. This optimist, therefore, not only systematically recalls positive information he has already received, he also has a propensity to better recall any good information he receives in the future. In other words, good information “sticks” to him while bad information “slides” off him. Symetrically, there would be a pessimistic steady state. To understand the local dynamics between these steady states, consider an optimist who encounters a long sequence of negative information. Their recency makes these bad events very memorable, and they form an association chain that can raise the recall probabilities of all bad events. Thus, a sequence of such events may push the individual to a pessimistic steady state. This sketch illustrates the possibilities of this approach.

Third, are the empirical predictions true in the data? The predictions with respect to λ espe-

cially permit out of sample tests, a point have repeatedly emphasized. Carrying out these tests is a natural next step, one that would begin the process of evaluating the practical relevance of the naive model outlined in this paper.

References

- Akerlof, George (1991). "Procrastination and Obedience," *American Economic Review*, 81(2), May, pp. 1-19.
- Anderson, R. C., Pichert, J. W., Goetz, E. T., Schallert, D. L., Stevens, K. V., and Trollip, S. R. (1976). "Instantiation of General Terms," *Journal of Verbal Learning and Verbal Behavior*, 15, 667-679.
- Banerjee, Abhijit (1992). "A Simple Model of Herd Behavior," *Quarterly Journal of Economics*, 107(3), 797-817.
- Barberis, Nick, Shleifer, Andrei and Vishny, Robert (1997). "A Model of Investor Sentiment," NBER Working Paper # 5926. Cambridge, MA: NBER.
- Bernard, Victor (1993). "Stock Price Reactions to Earnings Announcements: A Summary of Recent Anomalous Evidence and Possible Explanations," in Thaler (1993), pp. 303-40.
- Bartlett, F.C. (1932). *Remembering*. Cambridge: Cambridge University Press.
- Camerer, Colin (1995). "Individual Decision Making," in Kagel and Roth (1995).
- Camerer, Colin, Loewenstein, George, and Weber, Martin (1989). "The Curse of Knowledge in Economic Settings: An Experimental Analysis," *Journal of Political Economy*, 97(5), 1232-54.
- Campbell, John, and Mankiw, N. Gregory (1989). "Consumption, Income, and Interest Rates," in *NBER Macroeconomics Annual 1989*, ed. by Olivier Blanchard and Stanley Fischer. Cambridge, MA: MIT Press, 185-216.
- Campbell, John, and Shiller, Robert (1988a). "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," *Review of Financial Studies*, 1, 195-227.
- Campbell, John, and Shiller, Robert (1988b). "Stock Prices, Earnings and Expected Dividends," *Journal of Finance*, 43, 661-676.
- Caroll, Christopher, and Samwick, Andrew (1995). "The Nature of Precautionary Wealth," NBER Working Paper # 5193. Cambridge, MA: National Bureau of Economic Research.
- Conlisk, John (1996). "Why Bounded Rationality?" *Journal of Economic Literature*, 34(2), 669-700.
- Crovitz, H. F., and Schiffman, H. (1974). "Frequency of Episodic Memories as a Function of their Age," *Bulletin of the Psychonomic Society*, 4, pp. 517-518.
- Cutler, David, Poterba, James, and Summers, Lawrence (1989). "What Moves Stock Prices?" *Journal of Portfolio Management*, 15(3), 4-12.
- De Long, Bradford, Shleifer, Andrei, Summers, Lawrence, and Waldmann, Michael (1990). "Noise Trader Risk in Financial Markets," *Journal of Political Economy*, 98(4), 703-38.
- Deaton, Angus (1992). *Understanding Consumption*. Oxford: Oxford University Press.
- Dow, James (1991). "Search Decisions with Limited Memory," *Review of Economic Studies*, 58

- (1), January, pp. 1-14.
- Fischhoff, Baruch (1982). "For Those Condemned to Study the Past: Heuristics and Biases in Hindsight," in Kahneman, Slovic and Tversky (1982).
- Fudenberg, Drew and Levine, David (1997). *Theory of Learning in Games*, mimeo, Harvard University.
- Grether, David (1980). "Bayes Rule as a Descriptive Model: The Representativeness Heuristic," *Quarterly Journal of Economics*, 95 (3), November, 537-57.
- Hall, Robert and Mishkin, Frederic (1982). "The Sensitivity of Consumption to Transitory Income: Estimates from Panel Data on Households," *Econometrica*, 50, 461-81.
- Hayashi, Fumio (1985a). "The Effect of Liquidity Constraints on Consumption: A Cross-Sectional Analysis," *Quarterly Journal of Economics*, 100, 183-206.
- Hayashi, Fumio (1985b). "The Permanent Income Hypothesis and Consumption Durability," *Quarterly Journal of Economics*, 100, 1083-1113.
- Hamill, R., Wilson, T.D., and Nisbett, R.E. (1979). "Ignoring Sample Bias: Inferences about Collectivities from Atypical Cases," unblished manuscript, University of Michigan.
- Harvey, Andrew (1993). *Time Series Models*. 2nd. ed. Cambridge, MA: MIT Press.
- James, William (1890). *The Principles of Psychology*. Reprinted, Cambridge, MA: Harvard University Press (1983)
- Jappelli, Tullio and Pagano, Marco (1988). "Liquidity Constrained Households in an Italian Cross-Section," *Centre for Econ Policy Research*, discussion Paper: 257, August.
- Kagel, John, and Roth, Alvin, eds. (1995). *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.
- Kahneman, Daniel, and Tversky, Amos (1972). "Subjective Probability: A Judgement of Representativeness," *Cognitive Psychology*, 3, 430-54. Reprinted in Kahneman, Slovic, and Tversky (1982).
- Kahneman, Daniel, and Tversky, Amos (1973). "Availability: A Heuristic for Judging Frequency and Probability," *Cognitive Psychology*, 4, 207-32. Reprinted in Kahneman, Slovic, and Tversky (1982).
- Kahneman, Daniel, Slovic, Paul, and Tversky, Amos (1982). *Judgement Under Uncertainty: Heuristics and Biases*. New York, NY: Cambridge University Press.
- Kandel, Eric, Schwartz, James, and Jessell, Thomas, 3rd ed. (1991). *Principles of Neural Science*. New York, NY: Elsevier Science Publishing.
- Laibson, David (1997). "A Cue Theory of Consumption," mimeo, Harvard University.
- Lakonishok, Josef, Shleifer, Andrei and Vishny, Robert (1994). "Contrarian Investment, Extrapolation, and Risk," *Journal of Finance*, 49(5), pp. 1541-78.
- Lo, Andrew and MacKinlay, A. Craig (1990). "Data-Snooping Biases in Tests of Financial Asset Pricing Models," *Review of Financial Studies*, 3, 431-468

- Mackintosh, N.J. (1983). *Conditioning and Associative Learning*. Oxford: Clarendon Press.
- Mariger, Randall, and Shaw, Kathryn (1990). "Unanticipated Aggregate Disturbances and Tests of the Life-Cycle Model Using Panel Data," Board of Governors of the Federal Reserve, mimeo.
- Muth, John (1960). "Optimal Properties of Exponentially Weighted Forecasts," *Journal of the American Statistical Association*, 55 (290).
- Neisser, U. (1967). *Cognitive Psychology*. New York: Appleton-Century-Crofts.
- Nisbett, Richard, and Ross, Lee (1980). *Human Inference: Strategies and Shortcomings of Social Judgement*. Englewood Cliffs, NJ: Prentice-Hall.
- Pischke, Jörn-Steffen (1995). "Individual Income, Incomplete Information, and Aggregate Consumption," *Econometrica*, 63(4), July, pp. 805-840.
- Rabin, Matthew (1997). "Psychology and Economics," mimeo, University of California-Berkeley.
- Reder, Lynne, editor (1996). *Implicit memory and metacognition*, Carnegie Mellon Symposia on Cognition. Mahwah, N.J.: Lawrence Erlbaum.
- Roll, Richard (1984). "Orange Juice and Weather," *American Economic Review*, 74, 861-80.
- Ross, L., Lepper, M. R., and Hubbard, M. (1975). "Perseverance in Self Perception and Social Perception: Biased Attributional Processes in the Debriefing Paradigm," *Journal of Personality and Social Psychology*, 32, pp. 880-892.
- Rumelhart, David, McClelland, James, and the PDP Research Group (1986). *Parallel Distributed Processing : Explorations in the Microstructure of Cognition*. Cambridge, MA : MIT Press, 1986.
- Sargent, Thomas (1993). *Bounded Rationality in Macroeconomics*. Arne Ryde Memorial Lectures. Oxford and New York: Oxford University Press.
- Schacter, Daniel (1996). *Searching for Memory: The Brain, the Mind, and the Past*. New York, NY: Basic Books.
- Shiller, Robert (1989). *Market Volatility*. Cambridge, MA: MIT Press.
- Shleifer, Andrei, and Vishny, Robert (1997). "The Limits of Arbitrage," *Journal of Finance*, 52(1), 35-55.
- Stigler, George; Becker, Gary (1977). "De Gustibus Non Est Disputandum," *American Economic Review*, 67(2), 76-90.
- Thaler, Richard (1990). "Saving, Fungibility, and Mental Accounts," *Journal of Economic Perspectives*, 4(1) 193-205.
- Thaler, Richard (1993). *Advances in Behavioral Finance*. New York, NY: Russell Sage.
- Thompson, W. C., Reyes, R. M., and Bower, G. H. (1979). "Delayed Effects of Availability on Judgement". Manuscript, Stanford University.
- Trope, Yaacov (1978). "Inferences of Personal Characteristics on the Basis of Information Retrieved from One's Memory," *Journal of Personality and Social Psychology*, 36, 93-106. Reprinted in Kahneman, Slovic and Tversky (1982), 378-390.

- Tulving, E. and Schacter, D. L. (1990). "Priming and Human Memory Systems," *Science*, 247, 301-306.
- Tulving, E. and Thomson, D. M. (1973). "Encoding Specificity and Retrieval Processes in Episodic Memory," *Psychological Review*, 80, 352-373.
- Tversky, Amos, and Kahneman, Daniel, (1971). "Belief in the Law of Small Numbers," *Psychological Bulletin*, 2, 105-110. Reprinted in Kahneman, Slovic, and Tversky (1982).
- Waldfogel, Joel (1993), "The Deadweight Loss of Christmas," *American Economic Review*, 83(5), December pp. 1328-36.

Appendix

Lemma 1 *The optimal forecast satisfies:*

$$\begin{aligned}\hat{y}_t(h_t, e_t) &= x_t + \sum_{k=1}^{t-1} [w_{k,t}x_k + (1 \Leftrightarrow w_{k,t})(y_k \Leftrightarrow y_{k-1})] \\ \hat{\sigma}_t^2(h_t, e_t) &= \sigma_\nu^2 + \hat{\sigma}_{t-1}^2 \left(\frac{\sigma_\epsilon^2}{\hat{\sigma}_{t-1}^2 + \sigma_\epsilon^2} \right)\end{aligned}$$

where ns_t is the error to truth ratio: $\frac{\sigma_\epsilon^2}{\hat{\sigma}_t^2 + \sigma_\epsilon^2}$, and define: $w_{k,t} = \prod_{j=0}^{t-1} ns_{k+j}$. In the limit,

$$\begin{aligned}\lim_{t \rightarrow \infty} \hat{\sigma}_t^2 &= \sigma_*^2 = \frac{1}{2} \left(\sigma_\nu^2 + \sqrt{\sigma_\nu^2(\sigma_\nu^2 + 4\sigma_\epsilon^2)} \right) \\ \lim_{t \rightarrow \infty} w_{t(t+k)} &= \lambda^k\end{aligned}$$

where $\lambda = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_*^2}$,

Proof: Computing the optimal forecast is a straightforward application of the Kalman filter; see ch. 4, Harvey (1993).³⁶ Given the forecast rule, computing the steady requires setting $\hat{\sigma}_t^2 = \hat{\sigma}_{t+1}^2 = \sigma_*^2$:

$$\sigma_*^2 = \sigma_\nu^2 + \sigma_*^2 \left(\frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_\nu^2} \right)$$

Solving the resulting quadratic provides:

$$\sigma_*^2 = \frac{1}{2} \left(\sigma_\nu^2 + \sqrt{\sigma_\nu^2(\sigma_\nu^2 + 4\sigma_\epsilon^2)} \right)$$

As $t \rightarrow \infty$, $n_{kt} \rightarrow \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_*^2}$ meaning that $w_{kt} \rightarrow \lambda^{t-k}$. ■

Lemma 2 *Forgetting probabilities satisfy:*

$$E[f_{k(t+j)}x_k|e_t] = \begin{cases} 0 & \text{if } k > t \\ \frac{1 - \chi^{\nu(e_t)}}{1 - \rho} (1 \Leftrightarrow \rho^j)x_t & \text{if } k = t \\ \Leftrightarrow \rho^j \chi \mathcal{E}(e_t) & \text{if } k < t \end{cases}$$

Proof: When $k > t$, $f_{k(t+j)}x_k$ depends only on events at time greater than t . Independence across time, therefore, shows that $E[f_{k(t+j)}x_k|e_t] = 0$ in this case. When $k = t$, $E[f_{k(t+j)}x_k|e_t] = x_t E[f_{t(t+j)}|e_t]$. Breaking this apart:

$$E[f_{t(t+j)}|e_t] = E[\underline{f} \Leftrightarrow \chi a_{t(t+j)}|e_t] + \rho E[\underline{f} \Leftrightarrow \chi a_{t(t+j-1)}|e_t] + \dots + \rho^{j-1} E[\underline{f} \Leftrightarrow \chi a_{t(t+1)}|e_t]$$

³⁶A derivation for the steady state can be found in Muth (1960).

This equals $\frac{1-\rho^j}{1-\rho}(\underline{f} \Leftrightarrow \chi \mathcal{V}(e_t))$. Finally, when $k < t$, note that

$$\begin{aligned} E[x_k f_{k(t+j)}|e_t] &= E[x_k(\underline{f} \Leftrightarrow \chi a_{k(t+j)})|e_t] + \rho E[x_k(\underline{f} \Leftrightarrow \chi a_{k(t+j-1)})|e_t] + \dots \\ &\quad + \rho^j E[x_k(\underline{f} \Leftrightarrow \chi a_{kt})|e_t] + \dots + \rho^{t+j-k-1} E[x_k(\underline{f} \Leftrightarrow \chi a_{k(k+1)})] \end{aligned}$$

By independence, all terms here are zero except $\rho^j E[x_k(\underline{f} \Leftrightarrow \chi a_{kt})|e_t]$. Even here, $E[x_k \underline{f}|e_t] = 0$. This gives: $\Leftrightarrow \chi \rho^j E[a_{kt} x_k | e_t] = \Leftrightarrow \chi \rho^j \mathcal{E}(e_t)$. ■

Lemma 3 *Conditioning on e_t , time t beliefs satisfy:*

$$E[\hat{y}_t^R | e_t] = x_t + \chi \mathcal{E}(e_t) \frac{\lambda}{1 \Leftrightarrow \lambda} (1 \Leftrightarrow \lambda^{t-1})$$

Proof: Notice that $\hat{y}_t^R = \hat{y}_t \Leftrightarrow err_t^m$. This allows writing:

$$E[\hat{y}_t^R | e_t] = E[\hat{y}_t | e_t] \Leftrightarrow E[err_t^m | e_t]$$

Now, $E[\hat{y}_t | e_t] = x_t$. The second term can be written as:

$$\Leftrightarrow E[err_t^m | e_t] = \Leftrightarrow \sum_{k=1}^{t-1} \lambda^{t-k} E[x_k f_{kt} | e_t]$$

By Lemma 2, $E[x_k f_{kt} | e_t] = \Leftrightarrow \chi \mathcal{E}(e_t)$. Substitution provides that:

$$\begin{aligned} E[\hat{y}_t^R | e_t] &= x_t + \chi \mathcal{E}(e_t) (\lambda + \lambda^2 + \dots + \lambda^{t-1}) \\ &= x_t + \chi \mathcal{E}(e_t) \frac{\lambda}{1 \Leftrightarrow \lambda} (1 \Leftrightarrow \lambda^{t-1}) \end{aligned}$$

■

Lemma 4 *Conditioning on e_t , time $t+j$ beliefs satisfy:*

$$E[\hat{y}_{t+j}^R | e_t] = x_t \left(1 \Leftrightarrow \frac{\underline{f} \Leftrightarrow \chi \mathcal{V}(e_t)}{1 \Leftrightarrow \rho} (1 \Leftrightarrow \rho^j) \lambda^j \right) + (\rho \lambda)^j \chi \mathcal{E}(e_t) \frac{\lambda}{1 \Leftrightarrow \lambda} (1 \Leftrightarrow \lambda^{t-1})$$

Proof: Again, notice that $\hat{y}_{t+j}^R = \hat{y}_t \Leftrightarrow err_{t+j}^m$. The conditional expectation of the first term with respect to e_t equals x_t . The conditional expectation of the second term equals:

$$\Leftrightarrow E[err_{t+j}^m | e_t] = \Leftrightarrow \sum_{k=1}^{t+j-1} \lambda^{t+j-k} E[x_k f_{k(t+j)} | e_t]$$

Applying Lemma 2 tells us that the summands in this summation are zero for $k > t$. This leaves:

$$\Leftrightarrow \lambda^j x_t E[f_{kt} | e_t] \Leftrightarrow \lambda^j \sum_{k=1}^{t-1} \lambda^{t-k} E[x_k f_{k(t+j)} | e_t]$$

Again applying Lemma 2 to $E[x_k f_{k(t+j)}|e_t]$ provides:

$$\Leftrightarrow \lambda^j x_t E[f_{kt}|e_t] + \lambda^j \rho^j \chi \mathcal{E}(e_t) \sum_{k=1}^{t-1} \lambda^{t-k} = \Leftrightarrow \lambda^j x_t E[f_{kt}|e_t] + (\lambda \rho)^j \chi \mathcal{E}(e_t) \frac{\lambda}{1 \Leftrightarrow \lambda} (1 \Leftrightarrow \lambda^{t-1})$$

Putting these together:

$$\Leftrightarrow E[err_{t+j}^m|e_t] = x_t (1 \Leftrightarrow \lambda^j E[f_{k(t+j)}|e_t]) + (\lambda \rho)^j \chi \mathcal{E}(e_t) \frac{\lambda}{1 \Leftrightarrow \lambda} (1 \Leftrightarrow \lambda^{t-1})$$

Finally, Lemma 2, allows us to write: $E[f_{k(t+j)}|e_t] = \frac{f - \chi \mathcal{V}(e_t)}{1 - \rho} (1 \Leftrightarrow \rho^j)$. Substitution gives the stated formula:

$$E[\hat{y}_{t+j}^R|e_t] = x_t \left(1 \Leftrightarrow \frac{f \Leftrightarrow \chi \mathcal{V}(e_t)}{1 \Leftrightarrow \rho} (1 \Leftrightarrow \rho^j) \lambda^j \right) + (\rho \lambda)^j \chi \mathcal{E}(e_t) \frac{\lambda}{1 \Leftrightarrow \lambda} (1 \Leftrightarrow \lambda^{t-1})$$

■

Lemma 5 *The following are true:*

$$\begin{aligned} \text{sign}(E[x_k c(x \Leftrightarrow x_k)|x]) &= \text{sign}(x) \\ \text{sign}(E[n_k c(n \Leftrightarrow n_k)|n]) &= \text{sign}(n) \end{aligned}$$

Proof: I will show the first of the two equations, the proof for the second is exactly the same.

$$E[x_k c(x \Leftrightarrow x_k)] = \int_{-\infty}^{\infty} x_k c(x \Leftrightarrow x_k) dF_k$$

Breaking the integral at zero and applying symmetry of the x distribution gives:

$$\int_0^{\infty} x_k [c(x \Leftrightarrow x_k) \Leftrightarrow c(x + x_k)] dF_k$$

Since $x_k > 0$ in this equation, the sign of it equals the $\text{sign}(c(x \Leftrightarrow x_k) \Leftrightarrow c(x + x_k))$. Now,

$$c(x \Leftrightarrow x_k) \Leftrightarrow c(x + x_k) > 0$$

if and only if x is closer to x_k than to $\Leftrightarrow x_k$, which happens if and only if x is positive. Formally, since $c(\cdot)$ measures closeness, $c(x \Leftrightarrow x_k) > c(x + x_k)$ if and only if $|x \Leftrightarrow x_k| > |x + x_k|$. Squaring both sides, gives :

$$(x \Leftrightarrow x_k)^2 \Leftrightarrow (x + x_k)^2 > 0 \Leftrightarrow (2x)(2x_k) > 0$$

Since $x_k > 0$, this is equivalent to $x > 0$. This shows that:

$$\text{sign}(E[x_k c(x_t \Leftrightarrow x_k)]) = \text{sign}(x)$$

■

Lemma 6 *Associativeness implies:*

$$E[x_k x_t a_{kt}] > 0$$

Proof: Note that:

$$E[x_k x_t a_{kt}] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_k x_t c(x_k \Leftrightarrow x_t) dF_k dF_t$$

Breaking apart the integrals allows us to write:

$$\left(\int_0^{\infty} \int_0^{\infty} + \int_{-\infty}^0 \int_{-\infty}^0 + \int_0^{\infty} \int_{-\infty}^0 + \int_{-\infty}^0 \int_0^{\infty} \right) x_k x_t c(x_k \Leftrightarrow x_t) dF_k dF_t$$

Perform the integral transformation in the second and third integrals of $x_k \mapsto \Leftrightarrow x_k$ and $x_t \mapsto \Leftrightarrow x_t$. By symmetry of the F distribution and $c(\cdot)$, this becomes:

$$\begin{aligned} & 2 \left(\int_0^{\infty} \int_0^{\infty} + \int_0^{\infty} \int_{-\infty}^0 \right) x_k x_t c(x_k \Leftrightarrow x_t) dF_k dF_t = \\ & 2 \int_0^{\infty} \left(\int_0^{\infty} x_k c(x_k \Leftrightarrow x_t) dF_k + \int_{-\infty}^0 x_k c(x_k \Leftrightarrow x_t) dF_k \right) x_t dF_t \end{aligned}$$

Performing the transformation $x_k \mapsto \Leftrightarrow x_k$ now gives:

$$2 \int_0^{\infty} \int_0^{\infty} x_k [c(x_k \Leftrightarrow x_t) \Leftrightarrow c(x_k + x_t)] x_t dF_t$$

which as we saw in the previous proof is positive since for positive x_t , $c(x_k \Leftrightarrow x_t) > c(x_k + x_t)$. ■

Lemma 7 *Forecast errors satisfy:*

$$err_{t+1}^m = \rho \lambda err_t^m + \sum_{k=1}^t \lambda^{t+1-k} x_k (\underline{f} \Leftrightarrow \chi a_{k(t+1)})$$

Proof: Write:

$$err_{t+1}^m = \sum_{k=1}^t \lambda^{t+1-k} x_k f_{k(t+1)}$$

Using the fact that $f_{k(t+1)} = \rho f_{kt} + \underline{f} \Leftrightarrow \chi a_{k(t+1)}$, we get:

$$err_{t+1}^m = \rho \sum_{k=1}^{t-1} \lambda^{t+1-k} x_k f_{kt} + \sum_{k=1}^t \lambda^{t+1-k} (\underline{f} \Leftrightarrow \chi a_{k(t+1)})$$

Substituting in for err_t^m in the first term gives:

$$err_{t+1}^m = \rho \lambda err_t^m \sum_{k=1}^t \lambda^{t+1-k} (\underline{f} \Leftrightarrow \chi a_{k(t+1)})$$

completing the proof. ■

Lemma 8 *The variance, $Var[err_t^m|e_t]$ is less than $Var[err_t^i]$ for small χ and increases with χ*

Proof: Note that $Var[err_t^m|e_t]$ equals:

$$\sum_{k=1}^{t-1} \sum_{j=1}^{t-1} \lambda^{2t-k-j} E[x_k x_j f_{kt} f_{jt} | e_t]$$

When χ is close to zero, notice that the non-diagonal terms, where $k \neq j$ are also close to zero. To see, this notice that these terms equal:

$$\begin{aligned} E[(\underline{f} + \rho f_{k(t-1)} \Leftrightarrow \chi a_{kt})(\underline{f} + \rho f_{j(t-1)} \Leftrightarrow \chi a_{jt}) x_k x_j | e_t] &= \\ \rho^2 E[f_{k(t-1)} f_{j(t-1)} x_k x_j] + \chi^2 E[a_{kt} a_{jt} x_k x_j | e_t] &\approx 0 \end{aligned}$$

The last approximation is because the second term directly goes to zero as χ does, and the first term goes to zero since $f_{k(t-1)}$ and $f_{j(t-1)}$ only depend on each other through associativeness, as seen in Lemma 2. Since the non-diagonal terms are close to zero, let's focus on the diagonal terms:

$$E[(\underline{f} + \rho f_{k(t-1)} \Leftrightarrow \chi a_{kt})^2 x_k^2 | e_t]$$

Again, as χ goes to zero, this becomes a constant (as usual, taking $t \rightarrow \infty$), $\left(\frac{f}{1-\rho}\right)^2$ times x_k^2 . And, since $\frac{f}{1-\rho}$ is less than 1 this whole term will be less than $E[x_k^2]$. Therefore,

$$\begin{aligned} Var[err_t^i|e_t] &= \sum_{k=1}^{t-1} \lambda^{2t-2k} E[x_k^2] \\ &> \sum_{k=1}^{t-1} \lambda^{2t-2k} \left(\frac{f}{1 \Leftrightarrow \rho}\right)^2 E[x_k^2] \\ &\approx E[err_t^m|e_t] \end{aligned}$$

To see the increase with χ , first note that in the derivation above, as χ increases, the non-diagonal elements increase. Similarly, in the derivation of the diagonal elements, these also increase with χ . The sum of these terms, therefore, rises with χ .

Finally, the non-diagonal terms ($k \neq j$) illustrate an important phenomena. Consider the $Var[err_t^i|e_t] = \sum_{k=1}^{t-1} \lambda^{2t-2k} E[x_k^2]$. It contains no such cross-terms. They exist only because of associativeness. Lemma 6 tells us that the cross terms will be positive. This is intuitive: associativeness raises variance by systematically introducing a correlation in the recalled information. When these cross terms are sufficiently large (for example, as χ gets large), then $Var[err_t^m|e_t]$ may even be larger than $Var[err_t^i|e_t]$ ■