Massachusetts Institute of Technology
Department of Economics
Working Paper Series

A MEMORY BASED MODEL OF BOUNDED RATIONALITY

Sendhil Mullainathan

Working Paper 01-28
September 2000

Room E52-251
50 Memorial Drive
Cambridge, MA 02142

Massachusetts Institute of Technology
Department of Economics
Working Paper Series

A MEMORY BASED MODEL OF BOUNDED RATIONALITY

Sendhil Mullainathan

Working Paper 01-28
September 2000

Room E52-251
50 Memorial Drive
Cambridge, MA 02142

Massachusetts Institute of Technology
Department of Economics
Working Paper Series

A MEMORY BASED MODEL OF BOUNDED RATIONALITY

Sendhil Mullainathan

Working Paper 01-28
September 2000

Room E52-251
50 Memorial Drive
Cambridge, MA 02142

A Memory Based Model of Bounded Rationality

Sendhil Mullainathan*

September 20, 2000

Abstract

How do memory limitations affect economic behavior? I develop a model of memory grounded in psychology and biology research to investigate this question. Using this model, I study the case where people apply Bayes rule to the history they recall as if it were the true history. The resulting beliefs exhibit over-reaction on average. They also exhibit under-reaction with the model providing enough structure to allow predictions about which effect dominates when. I then apply this general framework to an otherwise standard model of consumption. It predicts the broad structure of consumption predictability as well as differences in marginal propensity to consume across different income streams. Most importantly, because it ties the extent of bias to a measurable aspect of the stochastic process being forecasted, the model makes novel, testable empirical predictions.

*Massachusetts Institute of Technology and National Bureau of Economic Research. e-mail: mullain@mit.edu. This paper was written while I was a graduate student at Harvard University. I am indebted to my thesis committee, Drew Fudenberg, Lawrence Katz and Andrei Shleifer for their generous advice and encouragement, and to Marianne Bertrand, Edward Glaeser, David Laibson, and Richard Thaler for many helpful discussions. I have also greatly benefitted from comments by three anonymous referees, George Baker, John Campbell, Caroline Hoxby, Erzo F.P. Luttmer, and participants at numerous seminars. Financial support from the Chiles Foundation is gratefully acknowledged.

1 Introduction

Memory limitations appear to affect the way we form forecasts. Consider purchasing a car. Some information used, such as the prices, may come from outside sources. But much will come from our recollections, everything from newspaper reports about a recall to recollections of particular friendly acts by a boss (perhaps suggesting a promotion and, therefore, the ability to buy a more expensive car). Despite its obvious importance in many facets of economic life, however, memory limitations are largely ignored.[1] In this paper, I attempt to develop a tractable model of human memory, one that has testable predictions about economic behavior.

To build this model, I cull two stylized facts from the broad research on memory by biologists and psychologists. 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.

While these facts describe the technology of memory, they don't tell us how memory is used. People may be naive and treat the histories they recall as true, and apply Bayes' rule to them. Alternatively, they may be sophisticated, by possessing complete knowledge of the recall technology and correct 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 and draws out the implications of the naive model.[2]

[1] See 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.

[2] I 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.

When memory is used in this way, several interesting features result. First, associativeness means that events affect beliefs not just through the information they convey, but also through the memories they evoke. Receiving a devestating referee report likely evokes many bad memories, other instances that erode self confidence. More generally, associativeness generates an over-reaction (on average) to information as each event draws forth similar, supporting, memories. Bad news cultivates pessimism by selectively evoking negative information. In a related vein, *completely uninformative* signals can influence beliefs if they affect what is recalled. Viewing a fictional speech by a laid-off worker on the difficulty of finding a decent job may affect beliefs because it may trigger other, more informative memories.

Second, because of rehearsal, the memories evoked by an event linger, causing errors in belief to persist. So the over-reaction caused by associativeness dissipates only slowly. More generally, events will continue to have an effect long after the information they contain has been discredited. 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.

These results—and others derived through similar reasoning—match many of the experimentally found biases in human inference, such as the greater effect of salient information, the hot hand or belief perseverance. Most interestingly, though, the model makes a set of out of sample predictions. It relates the extent of such phenomena to the stochastic process that individuals are forecasting. When the stochastic process requires little use of history (such as a random walk), there will be little use of memory and hence little of the biases described. This ability to relate the extent of the bais to a measurable aspect of the stochastic process being studied makes the model refutable and is formalized in Section 3.4.

To assess the effectiveness of the general model in economic contexts, I apply it to consumption within a simple Permanent Income Hyopthesis framework. Memory distortions here generate violations of the standard orthogonality predictions: consumption changes *can* be predicted using lagged

information. Intuitively, when individuals receive good news about their personal income, such as a glowing compliment from the boss, they are more likely to remember other good news, causing them to over-forecast their income. This generates predictability of future consumption changes. 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. Income sources will show a high marginal propensity to consume when memories play a large role in forecasting it, such as with personal labor income rather than with gains in one's 401(k). Most importantly, as in the general case, the model makes a set of out of sample predictions. It relates the extent of consumption predictability to a parameter of the income process that measures the importance of the history. Moreover, since income streams with a high MPC connote over-reaction, it further predicts that the income streams with the largest MPCs should also show the greatest negative correlation with future consumption. Section 4 discusses these predictions.

These results suggest that bounds on human memory can be fruitfully modeled. The results match psychological findings as well as empirical facts in consumption. It also generates out of sample 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 firm's 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, 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, income, obey 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. Assume that individuals have priors about the value of y which are normally distributed with mean $\hat{y}_0$ and variance $\hat{\sigma}_0^2$.

Each period with probability p an event e_t occurs, which will provide "soft" information about income. Each event e_t has an informative, x_t, and non-informative, n_t, component. When there is no event, I will write $e_t = (0, 0)$. Conditional on an event occurring, they are distributed:

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

where $E[x_t] = E[n_t] = 0$.[3] The covariance term, $\sigma_{xn} = cov(x, n)$, measures whether the neutral component typically appears with positive or negative information. A concrete example of an event might be hearing a friend describing his recent unemployment experience. The length of his

[3] In Mullainathan (1998), this assumption is discussed in greater detail. Normality implies that x_t and n_t are symmetrically distributed. Symmetry implies that neither good nor bad events are more memorable (in the sense of both evocativeness and vividness discussed below.

unemployment spell would be informative (x_t), while the fact that he has a pregnant wife with medical bills piling up would be uninformative (n_t).[4] The model includes uninformative, or neutral components, because they will affect recall probabilities.

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 the informative part of the event tells her something about the shock that period, its informativeness is incomplete and depends on σ_x^2/σ_z^2.

At the beginning of each period, she observes the event, e_t. She then combines this information with past information to form a forecast, $\hat{y}_t$. At the end of the period the true value of income is observed. The game is then repeated.

2.1.1 Perfect Memory Forecasts

The perfect memory forecast will serve as a useful base case against which one can compare the forgetful forecast. 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 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]

[4] Of 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.

[5] A 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.

[6] Contrast 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.

The optimal forecast can be easily derived using the Kalman Filter (see Harvey 1993). The posterior at time t will be normally distributed with mean $\hat{y}_t$ and variance $\hat{\sigma}_t^2$. In steady state beliefs these will equal (see Lemma 1 in the appendix):

$$\hat{y}_t(h_t, e_t) = x_t + \sum_{k=1}^{t-1} \left[\lambda^{t-k} x_k + (1-\lambda^{t-k})\Delta y_k\right] \quad (2)$$

$$\hat{\sigma}_t^2(h_t, e_t) = \sigma_*^2 \quad (3)$$

where $\Delta y_k = y_k - y_{k-1}$ and $\lambda = \frac{\sigma_\epsilon^2}{\sigma_*^2+\sigma_\epsilon^2}$.

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 $\Delta y_k = x_k + z_k + \epsilon_k - \epsilon_{k-1}$, that contributes $(1-\lambda^{t-k})x_k$. Summing shows that the total coefficient on x_k is unity. Second, Δy_k enters forecasts with weight $1-\lambda^{t-k} < 1$ as is clear from the formula. Third, y_k influences forecasts at $\lambda^{t-k-1}(1-\lambda) < 1$ 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. This is intuitive. When λ approaches zero, most of the variance comes from permanent shocks, and hence the process resembles a random walk. In this case, history matters the least and past values should receive the least weight.

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 $k < t$:

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

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 $h_t^R = (e_1^R, e_2^R, ..., e_{t-1}^R, y_1, ..., y_{t-1})$. 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:

$$e_k^R = \begin{cases} e_k & \text{with probability } r_{kt} \\ (0,0) & \text{with probability } 1 - r_{kt} \end{cases}$$

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 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"

[7]The 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.

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

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 much knowledge about memory.[9] I will focus on two of these features, which in my opinion are by far the most relevant ones for economists.

The first, *rehearsal*, states that recalling a memory increases future recall probabilities. Students quickly recognize this property: repetition strengthens memories. Experimental evidence on rehearsal can be found even at the neuronal level. Repeated firings between neurons strengthens their synaptic connections, or the strenght of the "memory" stored there.[10] At a more macro level, experiments with humans shows similar behavior. Two groups of subjects memorize the same list of words. One group then practices recall of this list periodically, while the other does not (both see it only once). After the same time has elapsed for both groups, the group that has been periodically recalling the list shows higher recall of the list. That these findings should seem so obvious is a testament to the intuitive appeal of the rehearsal assumption.

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. 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

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

[10]See Kandel, Schwartz and Jessell (1991) for a discussion. 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 target words is higher when the paired word is present.[11] 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.[12]

As these studies demonstrate, both rehearsal and associativeness have a strong experimental basis.[13] In fact, the most popular models of memory (and neural function generally), Parallel Distributed Processing Models, possess both features (Rumelhart and McClelland, 1986). Nevertheless, I do not mean to imply that these are the only "important" facts about memory, merely that these appear to be the most relevant to economists.[14]

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

[11] These 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 the effect is stronger when the words are connected.

[12] Laibson (1997) derives a theory of consumption based on preferences that exhibit a form of conditioning, which is related to associativeness (MacKintosh, 1983). 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).

[13] The evolutionary advantage of these two properties are easy to understand. Frequently encountered phenomena and memories similar to current circumstances are both more relevant. I have not formally pursued such intuitive notion to get at a more evolutionary or optimizing basis of memory. Such a model would require a precise understanding of the constraints on what memory mechanisms are even biologically feasible.

[14] 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 may not possess all the episodes that gave rise to them. Second, memory seems to be reconstructive in nature (Neisser 1967). 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. I ignore these 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.

time t, with $R_{(t-1)t} = 1$ and $r_{(t-1)t} = 1$. Note that $E[R_{kt}] = r_{kt}$. With this notation, we can write:

$$r_{kt} = \underline{m} + \rho R_{k(t-1)} + \chi a_{kt} \tag{4}$$

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 6 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 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 $c : (-\infty, \infty) \to (0, 1)$ be a closeness function (that is, an inverse distance function). Then similarity is defined as:

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

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: $a_{kt} = \frac{1}{2}\left(e^{-(x_t - x_k)^2} + e^{-(n_t - n_k)^2}\right)$. Thus $0 < a_{kt} < 1$ and $a_{tt} = 1$.

Substituting back in to the original equation provides:

$$r_{kt} = \underline{m} + \rho R_{k(t-1)} + \chi \frac{1}{2}\left(c(x_t - x_k) + c(n_t - n_k)\right) \tag{5}$$

and recall that I assume that $\underline{m} + \rho + \chi < 1$. It will also be useful to define the forgetting probability, $f_{kt} = 1 - r_{kt}$ and similarly $F_{kt} = 1 - R_{kt}$. Further define the constant $\underline{f} = 1 - \underline{m} - \rho$, so that we can write $f_{kt} = \underline{f} + \rho F_{k(t-1)} - \chi \frac{1}{2}\left(c(x_t - x_k) + c(n_t - n_k)\right)$.

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 Dynamics of Recall

Before moving to forecasts, it will be useful to see how recall works. Simple recursive substitution yields:

$$E[f_{kt}|e_k] = \left(\underline{f} - \chi E[a_{kt}|e_k]\right)\frac{1-\rho^{t-k}}{1-\rho} \tag{6}$$

$$\lim_{t\to\infty} E[f_{kt}|e_k] = \frac{\underline{f} - \chi E[a_{kt}|e_k]}{1-\rho} \tag{7}$$

for all $k < t$. 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.[15] Also, $E[a_{kt}|e_k]$ increases memorability. This term, which I will define as **vividness**, $\mathcal{V}(e_k)$, measures how strongly associativeness affects a memory. Memories that are very similar to a randomly drawn event will be more vivid. The are more likely to be triggered through associativeness and, therefore, more memorable.[16]

While vividness capture the strength of associations that an event possesses, it will also be useful to define the *average information* of these associated events. Define the **evocativeness** of event e_t is: $\mathcal{E}(e_t) = E[x_k a_{kt}|e_t]$. 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

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

[16] This 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.

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.

3.2 Limited Memory Expectations

We now turn to the forecasts of the forgetful individual. I will make a crucial assumption here: the forgetful individual applies the forecasting rule in equation 2 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 as: $\hat{y}_t^R(h_t^R, e_t) = \hat{y}_t(h_t^R, e_t)$ and $\hat{\sigma}_t^{2R}(h_t^R, e_t) = \hat{\sigma}_t^2(h_t^R, e_t)$.

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. 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 (see, for example, Reder 1996). Similarly, experiments have manipulated the memorability of information and tested whether individuals' decisions correct for this manipulation. Supportive of the naivete assumption, decisions are insensitive to this manipulation.[17] Of course, repetition and room for learning, may dramatically

[17] See, 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.

alter these findings. Nonetheless, the findings suggest that characterizing the naive decision maker would be a useful first step.[18]

Simple substitution gives the formula for the forgetful forecast:

$$\hat{y}_t^R(h_t^R, e_t) = x_t + \sum_{k=1}^{t-1} \left[R_{kt}\lambda^{t-k}x_k + (1-\lambda^{t-k})\Delta y_k \right] \quad (8)$$

$$\hat{\sigma}_t^{2R}(h_t^R, e_t) = \sigma_*^2 \quad (9)$$

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.

In order to contrast the perfect recall and forgetful forecasts, it will be useful to define a memory error. First, let $err_t = y_t - \hat{y}_t$ and $err_t^R = y_t - \hat{y}_t^R$ be the forecast errors of the perfect recall and forgetful forecasts respectively. Now define:

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

to be the *memory error*, that is the difference in forecasts caused by memory problems. Note that $err_t^R = err_t + err_t^m$, so that the memory error also measures how memory distorts the forecast error of the forgetful individual. With these definitions in hand, I now examine the determinants of beliefs.

Proposition 1 *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

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

[18] Preliminary 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.

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:

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

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 $\rho \to 0$, the effect disappears. We also see here that vividness now plays a role. As we saw in Proposition 6, 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. Vividnes, how associated a memory is, 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, the average information content of memories associated with an event, does influence beliefs contemporaneously. 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. An interesting implication is that even completely uninformative signals can affect beliefs.

Proposition 2 *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 $sign(\sigma_{xn} n_t)$.*

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

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

The first term is zero by symmetry of $\mathrm{c}(\cdot)$ 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:

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

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 $sign(\sigma_{xn} n_t)$ which establishes the first part. ■

The logic here is simple. Even though individuals disregard a signal with $x_t = 0$ as completely uninformative, their beliefs are still shaped by the memories these events trigger. The mediator in this process is σ_{xn}, which determines whether the neutral cue tends to appear with positive or negative information. For example, $\sigma_{xn} > 0$, a positive neutral cue ($n_t > 0$) selectively evokes other positive neutral cue ($n_k > 0$) memories. If $\sigma_{xn} > 0$, these memories will (on average) have $x_k > 0$ and hence the event is selecitvely evoking positive information memories.

These results relate to experimental findings that salient information has a greater effect on beliefs. Two experiments highlight the differential effect of evocativenss and vividness. 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.[19] 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.[20] 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 evocativeness more than vividness. 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 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

[19]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.

[20]A 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.

welfare." Another 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.

When the groups are asked to state their attitudes about welfare recipients, those receiving the story expressed far more unfavorable attitudes than a control group. Those 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.[21]

Together, these experiments illustrate the contrast between vividness and evocativeness. The inessential cues in the testimony (e.g. guacamole on the carpet) will not (on average) trigger other memories that condemn or exonerate the defendant. They do, however, make the testimony more vivid, making it more likely to be remembered and influence beliefs in the future. On the other hand, the welfare mother story will selectively trigger memories. Its evocativeness means that it will have greater contemporaneous impact.

[21] That 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. It is also worth noting that an 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.

3.3 Over-reaction and Under-reaction

The previous propositions illustrate how information content alone does not determine an event's impact; the memories it triggers also matters. But these propositions do not tell us about how forecast errors will be biased. 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 3 *Forecast errors are negatively correlated with the information in the latest event:*

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

The extent of this over-reaction increases with χ *and* λ*:* $\frac{\partial Cov(err_t^R, x_t)}{\partial \chi} < 0$ *and* $\frac{\partial Cov(err_t^R, x_t)}{\partial \lambda} < 0$.

Proof: Note that $err_t^R = err_t + err_t^m$ and that err_t is independent of x_t. Therefore, $Cov(err_t^R, x_t) = Cov(err_t^m, x_t)$. 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: $-\chi E[x_k x_t a_{kt}]$. 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:

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

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. ■

Intuitively, good information may lead to a rosier view of the past, which leads to forecasts that are too large, which leads to a negative forecast error.[22] Over-reaction increases as χ rises because

[22]Evidence on over-reaction 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

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. This is an extremely important point, which we will return to in Section 3.4.

The previous proposition 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 2 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:

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

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

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

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

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.

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 3.4.

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.[23]

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

Proposition 4 *Let $T > t$. When events are very memorable ($\underline{f}$ low, χ and ρ large), then*

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

Proof: (Sketch) I will present a proof for the case where $t \to \infty$ 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

[23]This result bears a little resemblance to the findings on belief perseverance (e.g. Ross, Lepper and Hubbard, 1975), the curse of knowledge (e.g. Camerer, Loewenstein and Weber, 1990) and on hindsight bias (Fischoff, 1982). Experiments in all three of these categories emmphasize the inability of subjects to "undo" information. See Mullainathan (1998) for a discussion.

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:

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

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

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

Substitution for the latter gives:

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

The third term can be written as:

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

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

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

Putting these terms together gives:

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

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 $\underline{f} \to 0$) 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^R_{t+1}. We know from Proposition 3 that x_t is negatively correlated to err^R_t. This implies a negative auto-correlation.

The results so far illustrate two conflicting forces: over- and under-reaction. 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. The following propositions quantify when belief changes are negatively (over-reaction) and when they are positively correlated (under-reaction) to lagged information.

Proposition 5 *Suppose that forgetting probabilities are small, so that $\underline{f}$ is high and ρ and χ are low. Then:*

$$Cov[\Delta\hat{y}^R_{t+1}, \Delta y_{t-1}] \;<\; 0 \tag{10}$$

When these probabilities are large, however:

$$Cov[\Delta\hat{y}^R_{t+1}, \Delta y_{t-1}] \;>\; 0 \tag{11}$$

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

$$\frac{\partial Cov[\Delta\hat{y}^R_{t+1}, \Delta y_{t-1}]}{\partial\lambda} \;<\; 0 \tag{12}$$

Proof: Now,

$$\Delta\hat{y}^R_{t+1} = \Delta\hat{y}_t - err^m_{t+1} + err^m_t$$

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

$$E[err^m_t \Delta y_{t-1}] - E[err^m_{t+1}\Delta y_{t-1}]$$

From Lemma 7, we can write err^m_{t+1} in terms of err^m_t. Substituting for this gives:

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

Reapplying Lemma 7 to err_t^m gives:

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

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

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

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

$$(1-\lambda\rho)\lambda\rho E[x_{t-1}err_{t-1}^m] + C\lambda(1-\lambda(1+\rho))$$

Substituting for the first part from the proof of Proposition 3 gives:

$$-\chi\frac{\lambda(1-\lambda^{t-1})}{1-\lambda}E[a_{kt}x_k x_t] + C\lambda(1-\lambda(1+\rho)) \tag{13}$$

Suppose events are very memorable, so that the forgetting probability, $\underline{f}$ is low and χ and ρ are high. Then $(\lambda(1-\lambda(1+\rho))C = (\lambda(1-\lambda(1+\rho))E[x_{t-1}^2(\underline{f}-\chi a_{(t-1)t})]$ 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:

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

which is the same as equation (13) except (i) it has been divided through by λ and (ii) $1-\lambda(1+\rho)$ has been replaced by $1-2\lambda(1+\rho)$. The first has no result on the sign and the second only makes it more negative since C is positive and $1-\lambda(1+\rho) > 1-2\lambda(1+\rho)$. Consequently if equation (13) 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 3.4. A greater emphasis on history implies over-reaction is larger and takes a longer time to undo.[24]

3.4 The Role of History

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:

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

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 ($t - k$ 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 thorough observations of y_t instead. This establishes who there will be slow learning. This

[24] Given both these over and under-reaction biases, it is natural to ask whether an individual might not be simply better off by simply ignoring their memory completely. Mullainathan (1998) shows that ignoring memory may reduce bias but almost surely raises variance (since information is lost).

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 who quickly a forgotten event can 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:

$$\begin{aligned} E[\hat{y}_t^R|e_t] &= x_t - \sum_{k=1}^{t-1} \lambda^{t-k} E[x_k f_{kt}|e_t] \\ & x_t + \sum_{k=1}^{t-1} \lambda^{t-k} (\chi E[x_k a_{kt}|e_t]) \\ & x_t + \chi \mathcal{E}(e_t) \left(\lambda + \lambda^2 + \lambda^3 + \cdots + \lambda^{t-1}\right) \end{aligned}$$

To get the second equation, we exploit the fact that $\underline{f} x_k$ 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 $t \to \infty$ for simplicity, the impact of selective recall is:

$$\chi \mathcal{E}(e_t) \left(\lambda + \lambda^2 + \lambda^3 + \cdots\right) = \chi \mathcal{E}(e_t) \frac{\lambda}{1-\lambda}$$

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.[25]

[25] Carroll and Samwick (1995) present a technique that allows us to estimate λ. Notice that $Var(\Delta^d y_t) = d\sigma_\nu^2 + 2\sigma_\epsilon^2$. By computing $\Delta_d y_t$ in the data for many different d and regressing them on d, one can back out the variances.

4 Application to Consumption

Having developed the general results, I now apply the model to the consumption decision of individuals.[26] 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. In the current model, this allows us to write consumption as:

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

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

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

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

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^0_{it} be the individual specific one, we write $y_{it} = y^0_{it} + \alpha_i \bar{y}_t$. 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.[27] Events are observed for both processes. Let $\bar{c}_t$ be aggregate consumption.

[26] Mullainathan (1998) presents an application to asset pricing as well.

[27] There 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.

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:

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

while at the aggregate level:

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

To see how this prediction works, note that:

$$Cov(\Delta c_{i(t+k)}, \Delta y_{it}) = E[\Delta \hat{y}^R_{i(t+k)} \Delta y_{it}] + E[err^m_{i(t+k-1)} \Delta y_{it}]$$

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

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

where because of independence, I have dropped terms such as $E[\Delta \hat{y}_{i(t+k)} \Delta \bar{y}_t]$. Applying Proposition 5, 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:

$$E[err^{0m}_{i(t+k-1)} \Delta y^0_{it}] + \alpha_i^2 E[\tilde{err}^m_{i(t+k-1)} \Delta \bar{y}_{it}]$$

where err_{it}^{0m} is the memory error for the idiosyncratic income component and $\bar{err}_{it}^{m}$ is the memory error for the aggregate component. Just as in the proof of Proposition 5, 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 $Cov(\Delta c_{i(t+k)}, \Delta y_{it}) < 0$. The partial with respect to λ_i come clearly from Proposition 5, 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:

$$Cov(\Delta \bar{c}_{(t+k)}, \Delta \bar{y}_t) = \bar{\alpha}^2 E[\Delta \hat{\bar{y}}^R_{(t+k)} \Delta \bar{y}_t] + \bar{\alpha}^2 E[\bar{err}^m_{i(t+k-1)} \Delta \bar{y}_t]$$

where $\bar{\alpha}$ is the average of α_i. Reapplying Proposition 5 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, though not always) found support in the literature (Hall and Mishkin 1982, Hayashi 1985a, 1985b, Jappelli and Pagano, 1988 and Mariger and Shaw, 1990). 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 1, 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:

$$MPC_s = Cov(\Delta c_t, \Delta y_{st}) = E[\Delta \hat{y}^R_{st} \Delta y_{st}] + E[err^m_{s(t-1)} \Delta y_{st}]$$

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

$$E[\Delta \hat{y}_{st} \Delta y_{st}] - E[\Delta err^m_{st} \Delta y_{st}]$$

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

$$E[err^m_{st} \Delta y_{st}] = \chi E[\mathcal{E}(e)x] \frac{\lambda}{1-\lambda}$$

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.

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. Testing these predictions will provide a way of refudiating this model. 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 two directions of future 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 18. 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.

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). "Sotck 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. Cambrige, 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, Micahel (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.

Fischoff, 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 Pagono, Marco (1988). "Liquidity Constrained Households in an Italian Cross-Section," *Centre for Econ Policy Research*, iscussion 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: Claredon 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:*

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

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,*

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

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

Proof: Computing the optimal forecast is a straightforward application of the Kalman filter; see ch. 4, Harvey (1993).[28] 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 $t \to \infty$, $n_{kt} \to \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_\nu^2}$ meaning that $w_{kt} \to \lambda^{t-k}$. ∎

Lemma 2 *Forgetting probabilities satisfy:*

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

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:

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

[28] A derivation for the steady state can be found in Muth (1960).

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

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

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

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

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

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

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

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

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

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

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

■

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

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

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

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

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

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

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

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

Putting these together:

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

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

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

■

Lemma 5 *The following are true:*

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

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

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

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

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

Since $x_k > 0$ in this equation, the sign of it equals the $sign(c(x-x_k) - c(x+x_k))$. Now,

$$c(x-x_k) - c(x+x_k) > 0$$

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

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

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

$$sign(E[x_k c(x_t - x_k)]]) = sign(x)$$

■

Lemma 6 *Associativeness implies:*

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

Proof: Note that:

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

Breaking apart the integrals allows us to write:

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

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

$$2 \left(\int_0^{\infty} \int_0^{\infty} + \int_0^{\infty} \int_{-\infty}^{0} \right) x_k x_t c(x_k - x_t) dF_k dF_t \quad =$$

$$2 \int_0^{\infty} \left(\int_0^{\infty} x_k c(x_k - x_t) dF_k + \int_{-\infty}^{0} x_k c(x_k - x_t) dF_k \right) x_t dF_t$$

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

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

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

Lemma 7 *Forecast errors satisfy:*

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

Proof: Write:

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

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

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

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

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

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:

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

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:

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

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

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

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 $Var[err_t^i|e_t]$ ■