IMPLICATIONS OF RATIONAL INATTENTION

CHRISTOPHER A. SIMS

ABSTRACT. A constraint that actions can depend on observations only through a communication channel with finite Shannon capacity is shown to be able to play a role very similar to that of a signal extraction problem or an adjustment cost in standard control problems. The resulting theory looks enough like familiar dynamic rational expectations theories to suggest that it might be useful and practical, while the implications for policy are different enough to be interesting.

I. INTRODUCTION

Keynes's seminal idea was to trace out the equilibrium implications of the hypothesis that markets did not function the way a seamless model of continuously optimizing agents, interacting in continuously clearing markets would suggest. His formal device, price "stickiness", is still controversial, but those critics of it who fault it for being inconsistent with the assumption of continuously optimizing agents interacting in continuously clearing markets miss the point. This is its appeal, not its weakness.

The influential competitors to Keynes's idea are those that provide us with some other description of the nature of the deviations from the seamless model that might account for important aspects of macroeconomic fluctuations. Lucas's 1973 classic "International Evidence..." paper uses the idea that agents may face a signal-extraction problem in distinguishing movements in the aggregate level of prices and wages from movements in the specific prices they encounter in transactions. Much of subsequent rational expectations macroeconomic modeling has relied on the more tractable device of assuming an "information delay", so that some kinds of aggregate data are observable to some agents only with a delay, though without error after the delay. The modern sticky-price literature provides stories about individual behavior that explain price stickiness and provide a handle for thinking about what determines dynamic behavior of prices.

Date: July 4, 2002.

Most recently, theories that postulate deviations from the assumption of rational, computationally unconstrained agents have drawn attention. One branch of such thinking is in the behavioral economics literature (Laibson, 1997; Bénabou and Tirole, 2001; Gul and Pesendorfer, 2001, e.g.), another in the learning literature (Sargent, 1993; Evans and Honkapohja, 2001, e.g.), another in the robust control literature (Giannoni, 1999; Hansen and Sargent, 2001; Onatski and Stock, 1999, e.g.). This paper suggests yet another direction for deviation from the seamless model, based on the idea that individual people have limited capacity for processing information.

That people have limited information-processing capacity should not be controversial. It accords with ordinary experience, as do the basic ideas of the behavioral, learning, and robust control literatures. The limited information-processing capacity idea is particularly appealing, though, for two reasons. It accounts for a wide range of observations with a relatively simple single mechanism. And, by exploiting ideas from the engineering theory of coding, it arrives at predictions that do not depend on the details of how information is processed.[1]

In this paper we work out formally the implications of adding information-processing constraints to the kind of dynamic programming problem that is used to model behavior in many current macroeconomic models. It turns out that doing so alters the behavior implied by these models in ways that seem to accord, along several dimensions, with observed macroeconomic behavior. It also suggest changes in the way we model the effects of a change in policy "rule" and in the way we construct welfare criteria in assessing macroeconomic policy. These aspects of the results are discussed in detail in later sections of the paper, which can be read independently of the mathematical detail in earlier sections.

II. Information Theory

The basic idea of information theory is to measure the rate of information flow as the rate of uncertainty-reduction. It therefore starts with a measure of uncertainty, called entropy. For a random variable X with pdf $p(X)$ the entropy is $-E[\log(p(X))]$. This formula applies whether the pdf is a density with respect to Lebesgue measure on R^k, so X is "continuously distributed", or with respect to a discrete measure on a finite or countable set of points, so that X is a discrete random variable. We think of $p \log p$ as zero for X values where $p = 0$. It is conventional to take the log in the

[1] Of course for psychologists, this may make the theory less interesting. But for economists, whose comparative advantage has been in using the optimization assumption to sweep aside psychological detail in modeling behavior, the ability of this information-theoretic approach to similarly sidestep psychological detail may be attractive.

formula to the base 2, so that the entropy of a discrete distribution with equal weight on two points is 1, and this unit of information is called one "bit".[2]

Then information is thought of as moving through a "channel", in which one enters input data, and output data emerges, possibly error-ridden. If we enter input data that is a realization of a random variable X with pdf p, and the channel provides as output a random variable Z whose pdf, conditional on X, is $q(Z \mid X)$, we can form a conditional pdf for $X \mid Z$ by Bayes' rule as

$$r(X \mid Z) = \frac{p(X)q(Z \mid X)}{\int p(x)q(Z \mid x)\, dx} \,. \tag{1}$$

Then the information acquired is the change in entropy,

$$E[\log_2(r(X \mid Z)) \mid Z] - E[\log_2(p(X))] \,. \tag{2}$$

This quantity need not be positive for every possible realized value of Z — we can see a Z that makes us more uncertain about X than we were before we saw Z — but on average across all values of Z, it is necessarily positive.

Here are two examples of transfers of information at the same rate. In the first example, we have a device (e.g. a telegraph key transmitting "dot" or "dash") that sends one of two signals, and does so without error. We can think of the signals as zeros and ones. If we draw a 0 or 1 from a distribution with equal probability on the two values, and transfer it according to this mechanism, then with each transmission we eliminate all uncertainty in the value of the draw, and thus transmit one bit of information.

In the second example, we draw a value for a continuously distributed $N(0,1)$ random variable and transfer it by a mechanism that contaminates it with an independent $N(0,1/3)$ noise. That is, $X \sim N(0,1)$, $\{Z \mid X\} \sim N(X,1/3)$, and therefore $\{X \mid Z\} \sim N(.75Z,.25)$. It is easy to verify that the entropy of a $N(\mu,\sigma^2)$ random variable is $\log_2 \sigma + .5(\log_2(2\pi e))$ and therefore that the information transmitted in this example by each observation on Z is $\log_2 1 - \log_2 .5 = \log_2 2 = 1$ bit.

Note that if this second channel had transmitted X without error, the rate of information transmittal would have been infinite. This corresponds to the fact that a real number can be represented as an infinite sequence of digits. If we could transmit real numbers without error, we could transmit arbitrary infinite sequences of integers

[2]Entropy is related to the Kullback-Leibler information $\mathcal{I}(p;q)$ in statistics (Schervish, 1995, p.115). For a random variable X with pdf p entropy is $-\mathcal{I}(p;\mathbf{1})$, where $\mathbf{1}$ is the (possibly improper) "flat" pdf that is identically one. That is, under the usual interpretation of Kullback-Leibler information, entropy is minus a measure of how hard it is to tell from observations on X that its pdf is p, not a uniform density.

(or of zeros and ones) without error, and thus send infinitely many bits in a single transmission.

The channel in our first example has a "capacity" of one bit per transmission, or time unit, say one second. This is the maximum rate at which the channel can transmit information.[3] We could use the channel inefficiently, for example by drawing the input zeros and ones from a distribution with probability of 0 $p \neq .5$. But if we actually had to send a sequence of 0's and 1's generated by i.i.d. draws from a distribution in which $p = .99$, say, we could in fact, by proper "coding" of the sequence, send it at as close as we like to a 1 bps rate rather than the $-.99 \log .99 - .01 \log .01 = .08$ bps rate that we would achieve by naively sending each 0 and 1 as it appeared.

One easily understood way to do this is with "block codes". For example, we could consider all 32 possible sequences of 5 zeros and ones. With our assumed p of .99, the sequence of 5 successive zeros is much more likely than any other sequence, so we could map that into a unit-length sequence of a single zero. The 5 sequences containing 4 zeros and 1 one are the next most likely, so we could map them to the sequences 111, 110, 1011, 1010 and 1000. All the remaining sequences of 5 zeros and ones would have to be mapped into sequences that begin with 1001, and thus would be as long or longer than their original length of 5. But since these remaining sequences are much less common in our signal sequence than the sequences consisting mostly of zeros, the *average* length of sequence we transmit will be well under 5.[4] Coding theorems in information theory show that methods like this can allow transmission at a rate approaching capacity through any channel, regardless of the distribution of the input that one wishes to send.

Notice that the coding theorem result implies that if we would like to be sending a $N(0, 1)$ signal with a $N(0, 1/3)$ error, as in our second example, but we have available only a discrete transmission device like the telegraph key of our second example, we can, by proper coding, send our normally distributed message with normally distributed error at 1 transmission per second using the error-free telegraph key.

The reader may wonder what kind of coding could accomplish this. Suppose $X \sim N(0, 1)$ and we code positive X to 0 and negative X to 1. This produces $\mathrm{Var}(X \mid Z) = .36$, approximately, slightly bigger than .25, which is our target optimal expected squared error. Of course the conditional distribution of $X \mid Z$ is a truncated

[3]Readers may recall, if they use modems with their computers, that modem speeds are rated in "bits per second", or bps. This is their channel capacity. Speed increases for modems plateaued at 56000 bps, because telephone connections are designed to have approximately that capacity, so no change in modem design can achieve greater speed.

[4]A detailed example like this is worked out in the appendix to my 1998 paper.

normal, and thus quite non-normal. More sophisticated codes, in general requiring that blocks of X values be coded jointly, could do better. There is no simple formula for optimal coding of continuously distributed sources into a discrete "alphabet", but there is a literature on methods to do it in particular physical settings (Gray and Neuhoff, 2000). The important point for our further discussion below, is that it can be done. Information transmission barriers that take the form of a restriction to sending discrete signals at a finite rate without error are not fundamentally different from restrictions to sending continuously distributed signals with contaminating noise.

With continuously distributed variables, some apparently natural specifications for channels can turn out to be unreasonable because they imply infinite capacity. For example, it can't be true that an actual channel is capable of transmitting an arbitrary real number X so that the output signal is $X+\varepsilon$, where ε is, say, $N(0,1)$, independent of X. If this were so, we could, by scaling up our input X to have arbitrarily large variance, achieve arbitrarily high transmission rates. The channel must be specified in such a way that there is no way, by adjusting input in the permissible range, to make the signal-to-noise ratio arbitrarily large. This point is central to our discussion below when we consider dynamically more complicated channels.

A potentially important point that we will ignore in what follows is that optimal coding generally introduces delay. We will consider dynamic optimization problems in which a decision must be made at each t on the basis of information at that date. We will assume that when the decision is made at t, data on X_s, $s \leq t$ has been observed via an optimally designed finite capacity channel whose output is available up through time t. If, say, the optimal channel requires Gaussian noise, but our available channel were actually a telegraph key, or a set of them, of the same capacity as the optimal channel, coding optimally will let us approximate the characteristics of the optimal channel arbitrarily well, but at the price of some delay. The excuses for ignoring the delay at this stage are (i) that the gap between the performance of finite-code-length systems and optimal systems declines exponentially in code length, so that coding delay is likely to be small in practice, and (ii) that when many sources of information are being coded simultaneously, it is possible to come close to transmission at the Shannon capacity rate without delay, even when the rate per source is low.[5]

[5]Conversion of a continuously distributed source to discrete form for transmission through a digital channel (like a telegraph key) is known as *quantization*. The theory is surveyed in Gray and Neuhoff (2000), with the results underlying the assertion (ii) discussed particularly in the section on vector quantization, II.B, p.291.

III. Optimization with information-flow constraints

Suppose we would like to minimize $E[(Y-X)^2]$, where Y is our action and X is a random variable that we must observe through a finite-capacity channel. What is the optimal choice for the conditional distribution of $Y \mid X$, subject to the requirement that the information flow about X required to generate Y is finite? Formally we would like, taking the pdf $p(\cdot)$ of X as given, to solve

$$\min_q \left\{ E[(Y-X)^2] = \int (y-x)^2 q(y \mid x) p(x)\, dy\, dx \right\} \tag{3}$$

s.t.

$$(\forall x) \int q(y \mid x)\, dx = 1 \tag{4}$$

$$-E\big[E[\log(q(Y \mid X)) \mid X]\big] + E\left[\log\left(\int q(x \mid Y) p(x) dx\right)\right] < C\,. \tag{5}$$

$$\tag{6}$$

This last expression is easy to interpret as the average reduction in entropy when we use observations on X to reduce uncertainty about Y, which seems to be the opposite of what we are interested in. However it turns out that the information flow is the same whether we are using observations on X to inform ourselves about Y or vice versa. The information flow between the two jointly distributed variables is called their mutual information.

When the X distribution is Gaussian, it is not too hard to show that the optimal form for q is also Gaussian and independent of X, so that Y and X end up jointly normally distributed.[6]

This example illustrates a more general point. The information-flow constraint results in an outcome that looks like a signal-extraction problem's outcome. The optimal behavior of Y is *as if* X were being observed with an i.i.d. error. The variance of Y is less than that of X, as is usual when actions have to be based on error-ridden data. But there are important deviations from signal-extraction results in the predictions of this approach for how response to X will change as the distribution of X is changed. For example, if the capacity constraint remains the same, doubling the

[6]When X has a given non-Gaussian marginal distribution, it is still optimal to make the distribution of $X \mid Y$ Gaussian and independent of Y, if possible. This can be accomplished by forming $\tilde{h} = \tilde{p}/\tilde{\varphi}$, where the $\tilde{}$ indicates Fourier transform (or characteristic function) and φ is the standard Normal pdf. Taking the inverse Fourier transform h of $\tilde{h}$ gives us a marginal pdf for y, which we can then multiply by a Gaussian conditional pdf for $X \mid Y$ to obtain a joint pdf. Finally we apply Bayes rule to the joint pdf to find $q(y \mid x)$. Of course for some p's (e.g. those with discontinuities, or discrete distributions), $\tilde{p}/\tilde{\varphi}$ is not square-integrable, so this method won't work.

variance of X will result in doubling the variance in the $Y \mid X$ distribution. Also, if $\text{Var}(Y \mid X) \leq \text{Var}(X)/3$, then if X starts being drawn from a distribution concentrated on two points, we expect the "error" in Y and the corresponding damping of response to completely disappear. The same 1 bps transmission rate is maintained with the error-free transmission of the two-point distribution as with the 3:1 Gaussian signal/noise ratio.

IV. Dynamic optimization with information-flow constraints: frequency domain

Suppose we observe a random vector Y and use it to inform ourselves about a random vector X that is jointly normally distributed with Y, i.e.

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(0, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}' & \Sigma_{yy} \end{bmatrix}\right) . \tag{7}$$

The entropy of a multivariate $N(\mu, \Omega)$ of dimension n is

$$-\frac{n}{2}(\log(2\pi) + 1) - \frac{1}{2}\log|\Omega| . \tag{8}$$

Therefore the reduction in entropy of X from observing Y (or vice versa) is

$$-\tfrac{1}{2}\log\left|I - \Sigma_{yy}^{-1}\Sigma_{xy}'\Sigma_{xx}^{-1}\Sigma_{xy}\right| \tag{9}$$

If Y and X are finite subsequences drawn from jointly stationary stochastic processes, then the Σ matrices appearing in (9) are Toeplitz forms and can be approximately diagonalized by Fourier transform matrices. If we do so, we can see that as the length of the two vectors gets longer, the information flow they represent converges to a rate per observation of

$$-\frac{1}{2\pi}\int_{-\pi}^{\pi}\log\left(1 - \frac{|S_{xy}(\omega)^2|}{S_x(\omega)S_y(\omega)}\right)d\omega , \tag{10}$$

where S_x and S_y are spectral densities for the Y and X processes and S_{xy} is their cross-spectral density.

This formula extends directly to continuous time, where it becomes (the only difference is the limits of integration)

$$-\frac{1}{2\pi}\int_{-\infty}^{\infty}\log\left(1 - \frac{|S_{xy}(\omega)|^2}{S_x(\omega)S_y(\omega)}\right)d\omega . \tag{11}$$

From (11) we can see one important point. In continuous time, the coherence of the X and Y processes must drop toward zero, at least on average, as $|\omega| \to \infty$. Otherwise the integral would not converge. This means that the power, or amount of variation, in the noise must grow large relative to that in the "signal" as we go to higher and higher frequency variation. Thus any action taken in response to the finite

information-flow-rate observations on X must be insensitive to the highest-frequency oscillations in X, meaning Y cannot respond sharply and quickly to X. Also, Y must itself show noisy idiosyncratic high-frequency randomness.

To see the implications of these principles, consider the simplest possible quadratic control problem: choosing Y to track X with loss $E(Y-X)^2$. Rather than the conventional constraints that Y is subject to adjustment costs or that Y is based on observation of X contaminated with exogenously specified noise, we assume instead that Y is chosen on the basis of an optimally coded transmission of data about X through a finite-capacity information channel.

If Y could be chosen on the basis of observations of future as well as past X, it would be easy to use frequency-domain methods to solve the optimal tracking problem. For the more realistic situation where Y_t can depend only on $\{X_s, s \leq t\}$, I have not been able find analytic formulas, but it is not too difficult to obtain example solutions using numerical methods. Formally, we solve the problem

$$\min_{b,c} E\left[(Y_t - X_t)^2\right] \tag{12}$$

subject to

$$X_t = \sum_{s=0}^{n} a_s \varepsilon_{t-s}\,, \qquad Y_t = \sum_{s=0}^{n} b_s \varepsilon_{t-s} + \sum_{s=0}^{n} c_s \nu_{t-s}\,, \tag{13}$$

$$\begin{aligned} -\tfrac{1}{2}\int_{-\pi}^{\pi} \log\left(1 - \frac{|\tilde{a}\tilde{b}|^2}{|\tilde{a}|^2(|\tilde{b}|^2 + |\tilde{c}|^2)}\right) d\omega \\ = -\tfrac{1}{2}\int_{-\pi}^{\pi} \log\left(1 - \frac{1}{1 + |\tilde{c}|^2/|\tilde{b}|^2}\right) d\omega < C\,, \end{aligned} \tag{14}$$

where the ε and ν processes are both i.i.d. $N(0,1)$ stochastic processes. Note that the left-hand side of (14) is just new notation for (11).

The solution for the case where a is a simple linearly declining sequence of weights, with $n = 31$, is displayed in Figure 1. Note that, as expected, b is smoother than a, with the smoothing distorting the match between a and b especially strongly near 0. In effect, the smoothing creates a delay in the reaction of Y to X. Also note that c is sharply peaked near zero, implying that high frequency variation in Y is dominated by noise. As indicated in the caption to the figure, in notation we use in all figures, this configuration represents data transfer at the rate .641 "bpt", which stands for "bits per time unit". If we though of the time unit as monthly (not unreasonable with 31 lags), implementation of this level of tracking would require less than one bit per month of information flow, a very modest requirement. This is possible in

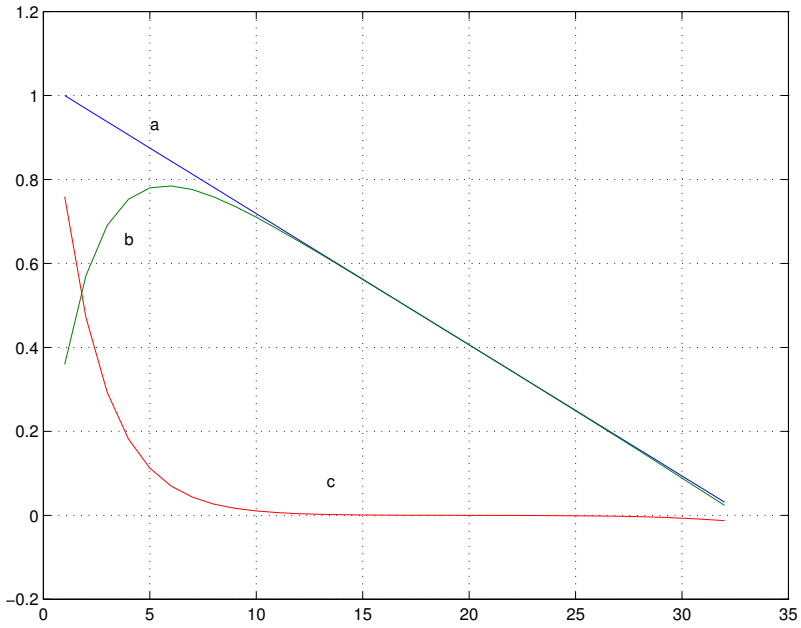

FIGURE 1. $C = .641$ bpt, $R^2 = .856$, linear a

this example because X, the variable being tracked, is modeled as highly serially correlated, so even a crude updating each month can track it fairly well.

If we allow an increased information flow, the tracking quickly becomes almost perfect, as in Figure 2.

If we instead tighten further the information constraint, as in Figure 3. we find the systematic part of Y smoother and more damped, and the noise in Y more serially correlated. Note that the absolute amount of high-frequency noise has gone down between Figures 1 and 3. This is because the capacity constraint relates only to the signal-noise ratio, $|\tilde{b}/\tilde{c}|$. When the information constraint becomes very tight, Y becomes nearly constant.

When the signal has less serial correlation, it has higher information content, so that a level of capacity that sufficed for quite good tracking in the examples we have examined above delivers much less accuracy. For example, if $a(s) = (.6^t - .5 \cdot .8^t) \cdot (1 - t/32)$, $s = 0, \ldots, 31$, then with a capacity of .71, close to that for Figure 1, we obtain Figure 4. Here the unconditional R^2 of Y and X is only .443, in contrast to the .856 achieved with a slightly lower information flow rate in the Figure 1 case. In fact the unconditional R^2 here is lower than that obtained in the Figure 3 case with a much lower information flow rate.

One final example for this section shows that the tendency of information constraints to induce delay is not confined to cases where a is discontinuous at 0. Here

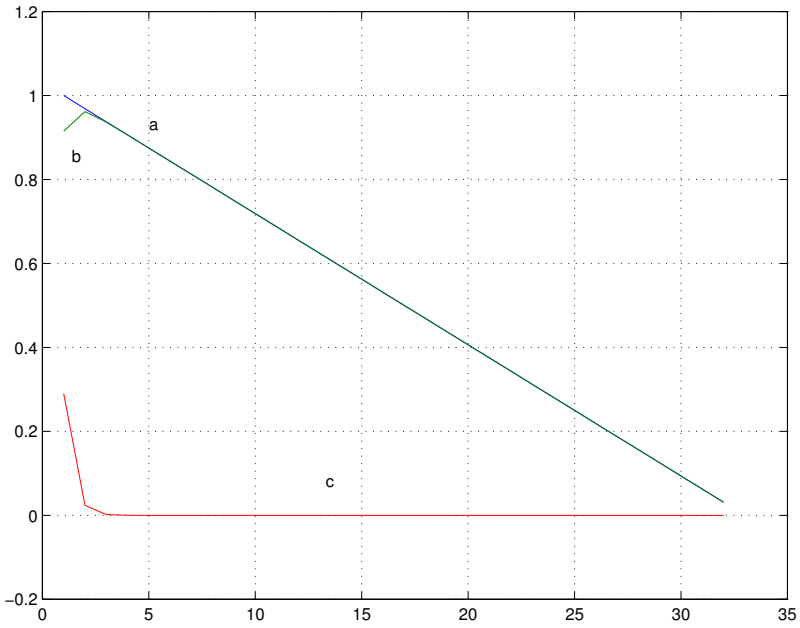

FIGURE 2. $C = 3.56$ bpt, $R^2 = .992$, linear a

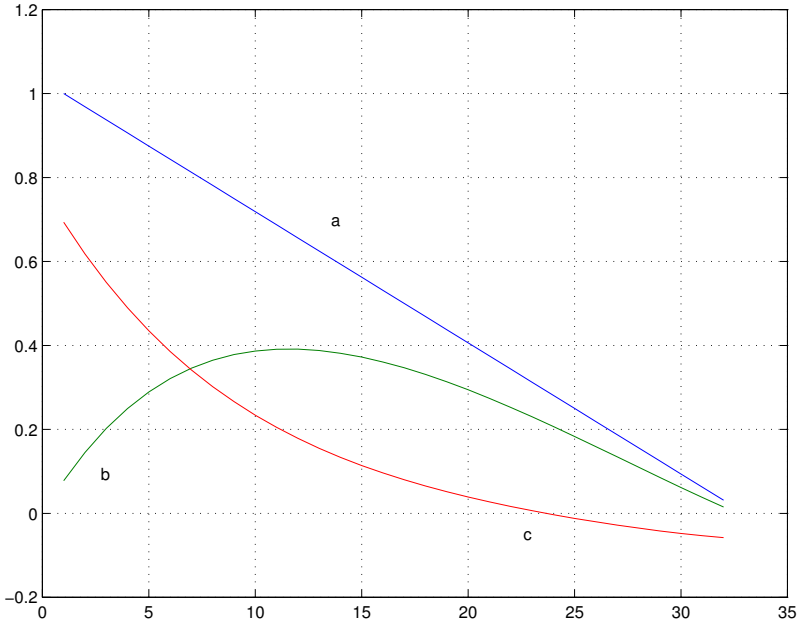

FIGURE 3. $C = .111$ bpt, $R^2 = .577$, linear a

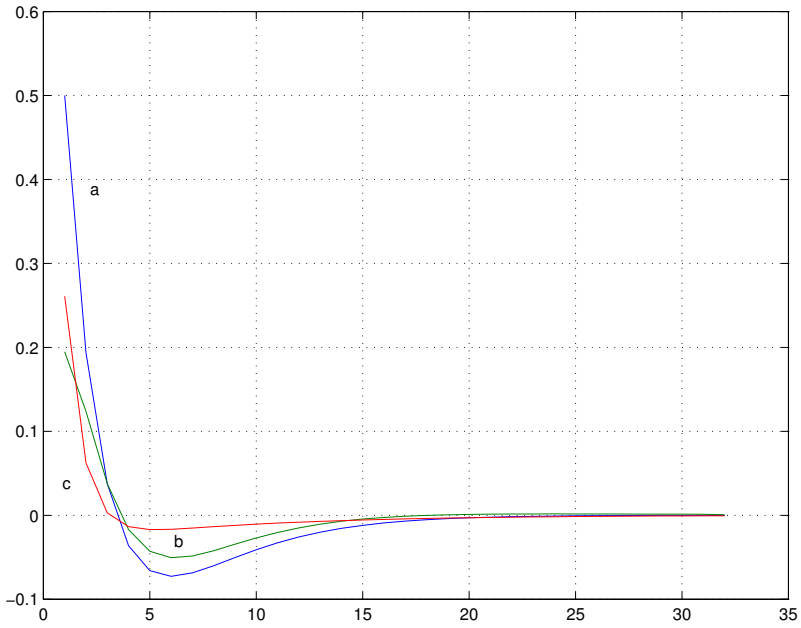

FIGURE 4. $C = .711$ bpt, $R^2 = .443$, rational whipsaw a

we set $a(s) = \sin(\pi(s+1)/32) \cdot (1 - s/32)$, $s = 0, \ldots, 31$ and capacity at .269. We emerge with an unconditional R^2 of .697, and, as can be seen from Figure 5, a distinct delay in the response of Y to X, as well as the expected damping.

From this set of examples we can confirm the idea that solutions to tracking problems with information-processing constraints mimic signal-extraction problems. But the information-processing approach makes the nature of the "noise" endogenous. We don't need to import the physicist's idea of i.i.d. experimental error or query ourselves about intuitively nebulous sources of exogenous noise in our view of economic data. Instead, the nature of the time series being tracked, together with the available information capacity, delivers a model for the noise.

If we take the tracking problems of this section as schematic representations of the way agents react to aggregate market signals, it may seem that the rates of information flow in the examples that show significant amounts of smoothing are implausibly low. A one bpt rate of information flow corresponds to making one yes-no decision per time period in reaction to the information flow. If the time unit is a month, it is clear that ordinary people's information-processing capacity is vastly higher than this. To justify information flow about major macroeconomic aggregate variables at this low rate, we must assume, as seems to be realistic, that most people devote nearly all of their information-processing capacity to types of information other than the time paths of macroeconomic variables. While it does seem realistic to suppose that people

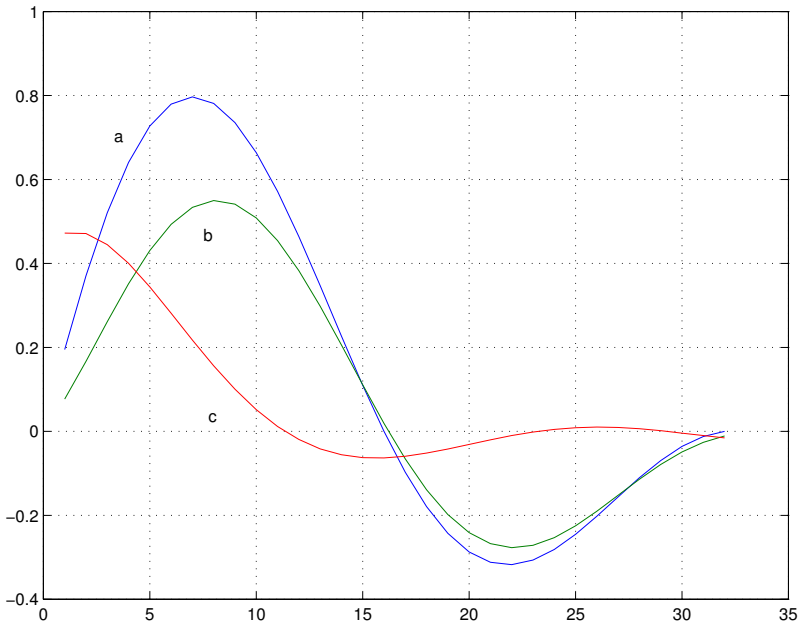

FIGURE 5. $C = .269$, $R^2 = .697$, oscillatory a

ordinarily pay only slight attention to aggregate economic time series, though, there is obviously plenty of room for them to shift more attention to these series' behavior when it becomes important to do so. A theory that treated capacity devoted to connecting aggregate data to economic behavior as exogenously fixed would be easier to produce, but it seems likely that treating this "economic capacity" as endogenous will be necessary to arrive at a realistic theory.

It would be interesting to consider a collection of tracking problems with a single capacity constraint. The amount of capacity allocated to each problem will then depend on its weight in the objective function and on the returns to capacity. Difficult tracking problems may get less capacity if they are difficult because of low marginal returns to additional capacity. With the same objective function and the same capacity limit, capacity allocations will change as the serial correlation and variance scales of the variables to be tracked change.

We can go some way toward these objectives by using the recursive formulation of the next section.

V. Linear-quadratic control with a capacity constraint and multivariate state

This section is technically dense. Readers may wish to read first (or only) the subsequent discussion of examples that apply this section's methods to economic

models. The single-state version of the permanent income problem discussed at the beginning of the next section brings out some of the main ideas with much simpler mathematics.

We consider a standard linear-quadratic control problem with imperfectly observed state, with the requirement that the observations on the state involve information transfer at a rate less than some given rate C. It is thus part of the optimization problem to choose the character of the observation error.

The problem is

$$\max_{C,S} E\left[\sum_{t=0}^{\infty} \beta^t(-S_t'AS_t - C_t'BC_t)\right] \tag{15}$$

subject to

$$S_t = G_0 + G_1S_{t-1} + G_2C_{t-1} + \varepsilon_t \tag{16}$$

$$S_t \,|\, \mathcal{I}_t \sim \mathcal{D}_t \tag{17}$$

$$S_{t-1} \,|\, \mathcal{I}_{t-1} \sim \mathcal{D}_{t-1} \tag{18}$$

and to the requirement that the rate of information flow at t implicit in the specification of $\mathcal{D}_{t-1}$ and $\mathcal{D}_t$ be less than C. Here S_t is the state at t, C_t is the control at t, and $\mathcal{I}_t$ is the information available at t. We are maintaining a convention (and an implicit assumption) that variables dated t are "known" at t, so that C_t and S_t are measurable with respect to the $\mathcal{I}_t$ information set.

Once the current period information has been received, the problem becomes a standard LQ control problem with measurement error, and we can apply certainty-equivalence. That is, if the (linear) decision rule that solves the deterministic version of the problem (that with $\mathrm{Var}(\varepsilon) = 0$) is

$$C_t = H_0 + H_1S_t\,, \tag{19}$$

then the optimal decision rule for the problem with measurement error is

$$C_t = H_0 + H_1\hat{S}_t\,, \tag{20}$$

where $\hat{S}_t = E[S_t \,|\, \mathcal{I}_t]$.

Let us now assume that both $\mathcal{D}_t$ and $\mathcal{D}_{t-1}$ are normal. Later we will return to consider to what extent we can deduce these properties of the signal-processing error from the structure of the problem. With $\mathcal{D}_t = N(\hat{S}_t, \Sigma_t)$ for each t, we will examine the situation where Σ_t is constant, verifying along the way that such a limiting case exists. If we let $\Omega = \mathrm{Var}(\varepsilon_t)$, then we can see from (16) that

$$\mathrm{Var}(S_t \,|\, \mathcal{I}_{t-1}) = \Psi = G_1'\Sigma G_1 + \Omega\,. \tag{21}$$

The information flow constraint can then be expressed as

$$-\log|\Sigma| + \log|\Psi| < 2\kappa\,, \tag{22}$$

where κ is channel capacity. In the case of a one-dimensional state, this constraint completes the characterization of the problem. When the state is multidimensional, however, (22) leaves open the possibility that $\Psi - \Sigma$ might not be positive semi-definite. This in effect implies that information flow can be kept low by "forgetting" some existing information, trading this off for increased precision about other dimensions of the state vector. Since transmission of information cannot produce this result, we need to add as a separate constraint

$$\Psi \succeq \Sigma\,, \tag{23}$$

where "$\succeq$" is interpreted as implying that the difference of left and right-hand sides is positive semi-definite.

We use the notation

$$V(S_t) = -S_t'\theta_2 S_t + \theta_1 S_t + \theta_0 \tag{24}$$

to denote the value function for the problem without capacity constraints and define

$$\hat{V}(\hat{S}_t) = E\Big[\sum_{s=0}^{\infty}\beta^t(-S_t'AS_t - C_t'BC_t)\;\Big|\;\mathcal{I}_t\Big]\,, \tag{25}$$

where the expectation is formed under the assumption that current and future C's are all being chosen subject to the capacity constraint on information flow.

Of course, since the information capacity constraint can't help in the optimization, we know that $E_tV(S_t) > \hat{V}(\hat{S}_t)$. (Here and henceforth we will use $E_t[\cdot]$ as shorthand for $E[\cdot\,|\,\mathcal{I}_t]$.) It turns out to be handy to characterize the optimization problem as being that of choosing Σ to minimize $E_tV(S_t) - \hat{V}(\hat{S}_t)$, in other words as that of choosing the structure of our uncertainty about the state so as to bring expected utility from the current date onward as close as possible to the expected value of the unconstrained value function. We begin by writing

$$\begin{aligned}E_tV(S_t) - \hat{V}(\hat{S}_t) =\\ -\operatorname{tr}\big((A + H_1'BH_1)\Sigma\big) + \beta E_t\Big[V(S^*_{t+1}) - V(S_{t+1}) + V(S_{t+1}) - \hat{V}(\hat{S}_{t+1})\Big]\,,\end{aligned} \tag{26}$$

where $S^*_{t+1} = G_0 + (G_1 + G_2H_1)S_t + \varepsilon_t$ is the value of S_{t+1} that would emerge if C_t were chosen optimally without uncertainty at t about the value of S_t. We use S_{t+1} to refer to the value of the state at $t+1$ when C_t is chosen subject to the capacity constraint, i.e. to satisfy (20). Then we write

$$\tilde{S}_{t+1} = S^*_{t+1} - S_{t+1} = G_2H_1(S_t - \hat{S}_t)\,. \tag{27}$$

Because of the LQ structure of the problem, the left-hand side of (26) will be a constant, determined entirely by variances, not the current estimate of the state. (Again, we'll verify this at the end.) Let this constant be M. Then we can write

$$(1-\beta)M = -\operatorname{tr}\big((A + H_1' B H_1)\Sigma\big) + \beta E_t\left[-\tfrac{1}{2}(\tilde{S}_{t+1}'\theta_2\tilde{S}_{t+1} + 2\tilde{S}_{t+1}'\theta_2 S_{t+1})\right] . \quad (28)$$

Using (27) and (16), we can then write the whole expression in terms of known matrices and Σ:

$$\begin{aligned}(1-\beta)M = & \\ & -\operatorname{tr}\Big(\big(A + H_1' B H_1 + \tfrac{1}{2}\beta(H_1'G_2'\theta_2 G_2 H_1 + H_1'G_2'\theta_2 G_1 + G_1'\theta_2 G_2 H_1)\big)\Sigma\Big) \\ & \qquad = -\operatorname{tr}(W\Sigma) . \quad (29)\end{aligned}$$

Our optimization problem then takes the form

$$\max_{\Sigma}\{\operatorname{tr}(W\Sigma)\} \quad (30)$$

subject to

$$-\log|\Sigma| + \log(G_1\Sigma G_1' + \Omega) \le 2\kappa \quad (31)$$

$$\Sigma \preceq G_1\Sigma G_1' + \Omega . \quad (32)$$

The information-theoretic constraints (31) and (32) have taken us out of the convenient linear-quadratic realm, but they are in some respects well-behaved. In particular, each defines a convex subset of Σ-space. To see that (31) defines a convex set of Σ's, observe that the second derivative of the left-hand side with respect to $\vec{\Sigma}$ is[7]

$$\Sigma^{-1}\otimes\Sigma^{-1} - (\Sigma + G_1^{-1\prime}\Omega G_1^{-1})^{-1}\otimes(\Sigma + G_1^{-1\prime}\Omega G_1^{-1})^{-1} , \quad (33)$$

where $\otimes$ is the Kronecker product. Obviously $\Sigma^{-1} \succeq (\Sigma + G_1^{-1\prime}\Omega G_1^{-1})^{-1}$, and the fact that for any square matrices X and Y, $X \succeq Y \Rightarrow X\otimes X \succeq Y\otimes Y$ then delivers the desired result. And that (32) defines a convex set of Σ's follows from its linearity in Σ and from the fact that $A \succeq 0$ and $B \succeq 0$ imply $A + B \succeq 0$.

Since our problem is therefore one with a linear objective function and convex constraint sets, it will have a uniquely defined maximized objective and we can characterize its solutions as maximizing the Lagrangian for certain values of the Lagrange multipliers on the constraints. One approach to solving this problem that I have found to work is to reparameterize in terms of the Choleski factor Φ of $\Lambda^* = \Psi - \Sigma$, where $\Phi'\Phi = \Lambda^*$ and Φ is upper triangular. This imposes the positive definiteness of $\Psi - \Sigma$, as required by (32), automatically. One can then maximize the Lagrangian

[7]We use the notation $\vec{\Sigma}$ to denote Σ stacked up, column by column, int a single vector. In deriving (33) and elsewhere we are using the identities $\overrightarrow{ABC} = (C'\otimes A)\vec{B}$ and $(d/d\Sigma)\log|\Sigma| = \Sigma^{-1}$.

with λ fixed. Note that the mapping from Λ^* to Σ, which must be evaluated in order to evaluate the objective function, is determined from the solution to

$$G_1 \Sigma G_1' + \Omega = \Sigma + \Lambda^* \,. \tag{34}$$

This is in the form of a standard discrete Lyapunov equation that can be solved,[8] e.g., with Matlab's `dlyap.m` program, or (if the equation is multiplied through by G_1^{-1}) by `lyap.m`.[9]

Once we have found the optimal Σ, we then find the corresponding $\mathrm{Var}(\xi_t) = \Lambda$ from the usual formula for the variance of a Gaussian distribution updated based on a linear observation:

$$\Sigma = \Psi - \Psi(\Psi + \Lambda)^{-1}\Psi \,. \tag{35}$$

Equation (35) can be solved for Λ^{-1}, yielding

$$\Lambda^{-1} = \Sigma^{-1} - \Psi^{-1} \,. \tag{36}$$

Note that Λ^{-1}, like Λ^*, is likely to be near singular. When this occurs, it means that it is efficient to observe only a certain linear combination (or combinations) of the state variables rather than the whole state vector. Measurement error on the other linear combinations is effectively infinite.

With the problem solved, we can form a dynamic system that characterizes the evolution of the vector $[S_t, \hat{S}_t, C_t]$ as it is driven by the underlying structural shocks

[8] If G_1 has a unit eigenvalue or pairs of eigenvalues that are reciprocals of each other, this equation can't be solved for Σ. This means for example that in the permanent income examples below, the algorithm fails if the exogenous income process has a unit root. This does not mean the problem is unsolvable, or even especially difficult in principle. But further cleverness will be required to produce a parameterization that works automatically in these cases. Also, some values for the Λ^* matrix may imply that there is no corresponding $\Sigma \succeq 0$. Numerical routines that search over Φ of course must take this into account, but this will not create important difficulties so long as $|\Sigma| > 0$ at the solution. Also, finding a feasible starting value for Λ^* may be a problem. If Ω is nonsingular, the solution to $\Lambda^* = G_1 \Sigma G_1' + \Omega - \Sigma$ is always p.s.d. if Σ is taken small enough.

[9] These routines are part of an extra-cost Matlab toolbox. A slower but slightly more general program called `lyapcs.m` is available on my web site. `lyap.m` quits without producing output under certain conditions in which there is a solution, but it is not unique. `lyapcs.m` attempts to provide a solution in these conditions. Because G_1 generally has both stable and unstable eigenvalues, doubling algorithms do not avoid the need for an eigenvalue decomposition of G_1 in this problem and thus lose their main advantage over the Schur-decomposition-based methods in the programs cited here.

ε_t and the information-processing-induced measurement error ξ_t. The equations are

$$S_t = G_0 + G_1 S_{t-1} + G_2 C_{t-1} + \varepsilon_t \tag{16}$$

$$\hat{S}_t = (I - \Sigma\Lambda^{-1})(G_1 + G_2 H_1)\hat{S}_{t-1} + \Sigma\Lambda^{-1}(S_t + \xi_t) \tag{37}$$

$$C_t = H_0 + H_1 \hat{S}_t \, . \tag{20}$$

This system can then of course be used to generate impulse responses in the usual way.

To justify the assumptions we have made along the way, it must turn out that the system formed by (16), (37) and (20) is consistent with stationarity of the $[S_t, \hat{S}_t, C_t]$ process. With stationary disturbance terms, we need, therefore, that the matrix

$$\begin{bmatrix} I & 0 & 0 \\ -\Sigma\Lambda^{-1} & I & 0 \\ 0 & -H_1 & I \end{bmatrix}^{-1} \begin{bmatrix} G_1 & 0 & G_2 \\ 0 & (I - \Sigma\Lambda^{-1})(G_1 + G_2 H_1) & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{38}$$

has all its roots less than one in absolute value.

It can be shown that the eigenvalues of this matrix are those of the two smaller matrices $G_3 = G_1 + G_2 H_1$ and $\Sigma\Psi^{-1}G_1$. The first of these, G_3, is the matrix characterizing the dynamics of the state in the model with no information capacity constraint. It will therefore have all its eigenvalues less than one in absolute value whenever the problem with no capacity constraint has a stable solution. The second of the two matrices seems likely also always to have all its values less than one in absolute value, whenever we can solve for Ψ and Σ, and I have yet to encounter an example where it has eigenvalues larger than one, but proving this seems to be challenging.

If κ is set too low, it is possible that the intersection of the sets of Σ's defined by (22) and (23), with Ψ defined by (21), is empty, in which case no stationary solution satisfying the capacity constraint is possible. This makes intuitive sense. In most rational expectations problems G_1 has at least some unstable roots, so that if C were never changed, S would blow up exponentially. Specification of too small a κ may in effect prevent C from moving enough, or accurately enough, to prevent this explosive behavior. But there is always a solution if κ is large enough.

VI. A Rational Inattention Version of Permanent Income Theory

We consider a standard linear-quadratic (LQ) permanent income example, in which an agent maximizes

$$E\left[\sum_{t=0}^{\infty} \beta^t (C_t - .5C_t^2)\right] \tag{39}$$

subject to

$$W_t = R \cdot (W_{t-1} - C_{t-1}) + Y_t \,, \tag{40}$$

where as usual we interpret W as wealth, C as consumption, Y as labor or endowment income, and R as the gross interest rate. We will suppose that the agent devotes limited capacity to observing Y and W, so that C_t is always being chosen as if Y and W up to the current date are known only up to some pattern of random error. We will use the idea of a finite Shannon capacity for the agent to derive the form of the pattern of random error and the effects of it on the way C and W respond to Y. This problem fits precisely into the framework of the general LQ control problem with capacity constraint that we considered in the previous section.

To begin with the simplest possible case, suppose Y_t is i.i.d. with mean $\bar{Y}$ and $\beta R = 1$. Then W_t by itself can serve as state variable. This model fits in to the general framework of the previous section, but it is much simpler to solve because of the one-dimensional state. Aspects of the problem that require numerical optimization in the multivariate case here can be solved analytically.

The optimal decision rule for the deterministic problem is simply

$$C_t = (1-\beta)W_t + \beta\bar{Y} \,. \tag{41}$$

Optimal policy with finite capacity will therefore have the form

$$C_t = (1-\beta)\hat{W}_t + \beta\bar{Y} \,. \tag{42}$$

To fit this problem to the notation of the previous section, we would make the notational correspondences $S_t \sim W_t$, $C_t \sim C_t - 1$[10], $G_0 \sim \bar{Y} + R$, $G1 \sim R$, and $G2 \sim -R$.

Rather than simply invoke the more general framework, we proceed here to work out this example analytically, as this may aid understanding. With a quadratic loss function, minimization of losses subject to an information-flow constraint implies that the conditional distribution of W_t given information at t will be $N(\hat{W}_t, \sigma_t^2)$.[11] Then the budget constraint (40) implies that

$$E_t[W_{t+1}] = \hat{W}_t \tag{43}$$

$$\mathrm{Var}_t[W_{t+1}] = R^2\sigma_t^2 + \omega^2 \,, \tag{44}$$

where ω^2 is the variance of Y_t. To keep things linear, we assume now that the distribution of Y_t is Gaussian. With a finite capacity κ per time unit, then, the

[10]The general model has no linear term in the objective function, so we have to recenter C at its satiation level of 1 in order to match the general setup's notation. This entails adjusting the constant term G_0 in the budget constraint also.

[11]This follows because the entropy of a pdf with a given variance σ_t^2 is maximized when the pdf is Gaussian.

optimizing agent will choose a signal that reduces the conditional standard deviation of W_{t+1} by

$$\kappa = \tfrac{1}{2}\big(\log(R^2\sigma_t^2 + \omega^2) - \log(\sigma_{t+1}^2)\big) \,. \tag{45}$$

This equation has a steady state at

$$\bar{\sigma}^2 = \frac{\omega^2}{e^{2\kappa} - R^2} \,. \tag{46}$$

In steady state the agent behaves as if observing a noisy measurement $W_{t+1}^* = W_{t+1} + \xi_{t+1}$, with ξ_{t+1} independent of all previous random disturbances to the system and with

$$\mathrm{Var}(\xi_{t+1}) = \frac{\bar{\sigma}^2(R^2\bar{\sigma}^2 + \omega^2)}{(R^2-1)\bar{\sigma}^2 + \omega^2} \,. \tag{47}$$

We compute (47) by setting the conditional variance of W_{t+1} given W_{t+1}^* and information at t to be $\bar{\sigma}^2$.

In some respects this model reproduces results of standard permanent income theory for this special case. Consumption and estimated wealth are random walks, in the sense that $E[C_{t+s} \mid \{C_v, v \le t\}] = C_t$ and $E[\hat{W}_{t+s} \mid \{\hat{W}_v, v \le t\}] = \hat{W}_t$. However consumption is not a martingale, meaning that it is not true that $E_t C_{t+s} = C_t$, where E_t is expectation conditional on *all* information available at t. That is, variables other than C's own past, and wealth and income variables in particular, may help predict future values of C.

To get a full solution for the system in terms of the exogenous process Y and the information-processing error ξ, we can assemble the budget constraint (40), the decision rule (42), and the regression formula describing how $\hat{W}$ is revised based on noisy observation:

$$\hat{W}_t = \hat{W}_{t-1} + \theta(W_t + \xi_t - \hat{W}_{t-1}) = (1-\theta)\hat{W}_{t-1} + \theta W_t + \theta\xi_t \,. \tag{48}$$

These form the system of difference equations

$$\begin{bmatrix} 1 & 0 & 0 \\ -\theta & 1 & 0 \\ 0 & -1+\beta & 1 \end{bmatrix} \begin{bmatrix} W_t \\ \hat{W}_t \\ C_t \end{bmatrix} = \begin{bmatrix} R & 0 & -R \\ 0 & 1-\theta & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} W_{t-1} \\ \hat{W}_{t-1} \\ C_{t-1} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Y_t - \bar{Y} \\ \xi_t \end{bmatrix} \,. \tag{49}$$

The classic permanent income theory literature characterizes the distributed lag relationship between C and Y, emphasizing that C should respond only slightly to temporary changes in Y, which are the only kind that occur in this model with no serial correlation. In fact, in this model without information constraints

$$C_t = (1-\beta)\left(W_0 + \sum_{s=1}^{t}(Y_s - \bar{Y})\right) + \beta\bar{Y} \,. \tag{50}$$

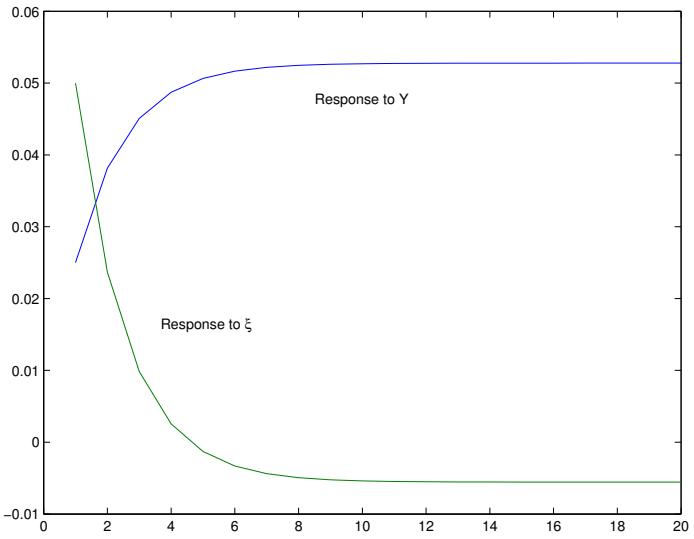

FIGURE 6. Responses of C in the Permanent Income Model with Serially Uncorrelated Y

This implies of course that the impulse response of C to Y shocks is flat, with an immediate upward jump that then persists indefinitely. When we solve the system (49), we get instead a response to Y that is "rounded off", with a steady rise to a flat asymptote that is above the response for the case without capacity constraint. The delay is of course because of the need to separate signal from noise, and the fact that the asymptote is above that for the standard case reflects the fact that an income shock, initially undetected, accumulates interest for a while before consumption responds fully. The information-processing error is much less serially correlated than the part of C that responds to Y. The two impulse responses — to $Y_t - \bar{Y}$ and to ξ_t — are shown in Figure 6. The figure assumes a discount factor of .95 and a regression coefficient θ of $\hat{W}_t$ on the observed $W_t + \xi_t$ of .5. This in turn implies that channel capacity is $-\frac{1}{2}\log_2(1-\theta^2) = .21$.

The noticeable smoothing and delay of response to income shock in this example seems to rest on assuming a very low bit rate of information transmission. While this rate is clearly well below the total information processing capacity of human beings, it may not be unreasonable as the rate assigned to processing economic data, particularly aggregate economic data, when there are many competing demands on capacity.

For example, we can consider what happens when the income process contains several components. We will suppose that it has two components with high innovation

variance and modest serial correlation, and another that has much lower innovation variance and stronger serial correlation. We might think of the latter as an "aggregate" component of the income of an individual whose personal fortunes are dependent, month to month, much more on local and idiosyncratic uncertainties than on aggregate economic fluctuations.

The setup here will be exactly as in the previous example, except for the specification of the exogenous Y process. Here we assume

$$Y_t = \bar{Y} + X_t + Z_t + \varepsilon_{Yt} \tag{51}$$

$$X_t = .97X_{t-1} + \varepsilon_{Xt} \tag{52}$$

$$Z_t = .9Z_{t-1} + \varepsilon_{Zt}\,, \tag{53}$$

with

$$\text{Var}\begin{bmatrix}\varepsilon_{Yt}\\ \varepsilon_{Xt}\\ \varepsilon_{Zt}\end{bmatrix} = \begin{bmatrix}.01 & 0 & 0\\ 0 & .0001 & 0\\ 0 & 0 & .003\end{bmatrix}\,. \tag{54}$$

So the X component of income is to be thought of as the stable, slow-moving "aggregate". This model again falls in to our general framework, but this time cannot be solved by hand. Figure 7 shows the responses to the three true component income shocks and the three error shocks, with channel capacity .72 bpt, four times that in Figure 6. Each of the responses to true components is accompanied by a horizontal line showing the level of the flat optimal response in the absence of capacity constraints. The response to Z, the large-variance serially correlated component, is similar in shape to the response to the single shock in the single-state example. The response to the serially uncorrelated component is essentially undamped, while the response to the low-shock-variance, near-unit-root component is extremely damped, with the response rising steadily throughout the 20 periods plotted.

If we triple channel capacity, to about 2.1 bpt, we get the responses shown in Figure 8. Here the response to the X component shock has almost the same form as the response to the single i.i.d. shock in the one-state example, but total channel capacity is ten times greater. In these multi-state examples the optimal use of channel capacity is to track closely the more unpredictable components of labor income, allocating proportionately much more observation error to the slower-moving "aggregate" component. Of course in reality people have many more than three economic time series to monitor and much more to keep track of than just the economic side of their lives. Thus it is not unreasonable to model them as reacting to macroeconomic data as if it reached them through a low-capacity channel.

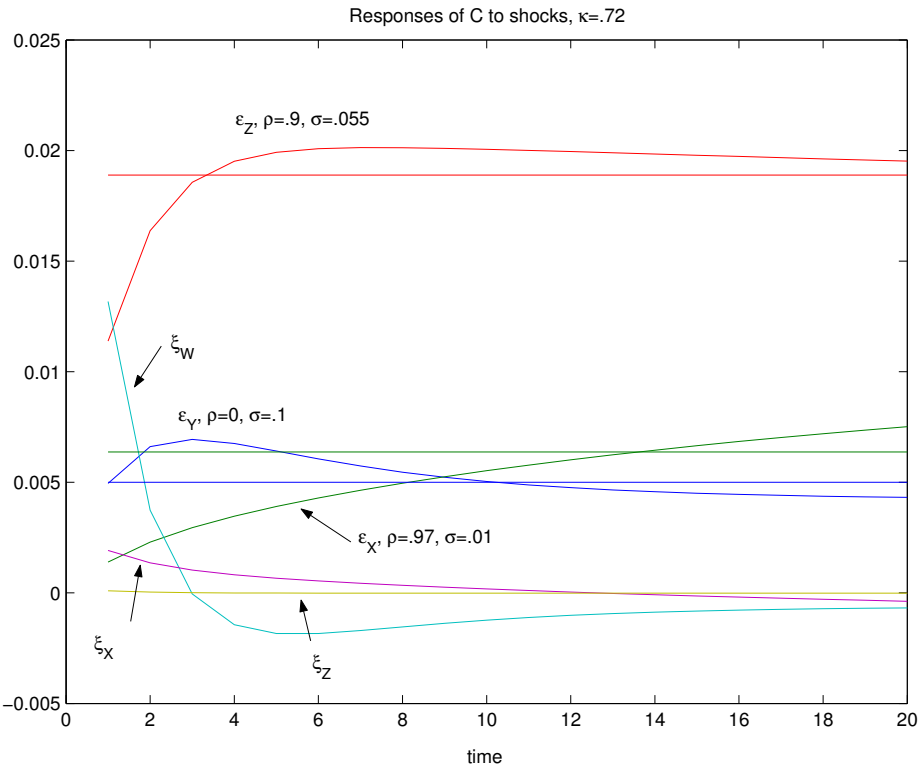

FIGURE 7

VII. IMPLICATIONS FOR MACROECONOMIC TIME SERIES BEHAVIOR

While the examples worked out above are far from constituting a general equilibrium theory based on these ideas, they suggest likely qualitative behavior of such a full equilibrium model. This way of thinking can provide a basis for critical thinking about the results of other approaches before it has been developed to the point of providing a fully worked out alternative.

As I noted in a previous paper (1998), it is apparent from nearly any VAR study and from the bivariate investigations of Keating (1997) that most cross-variable relationships among macroeconomic time series are smooth and delayed, as would be predicted by an information-constraint approach. It is possible (as is also shown in the 1998 paper) to reproduce this behavior with conventional DSGE models by introducing adjustment cost mechanisms pervasively, for both prices and quantities. The required adjustment costs are large, however, and the dynamics of the model starts to depend mainly on these mechanisms, which are hard to interpret in terms of microeconomic common sense.

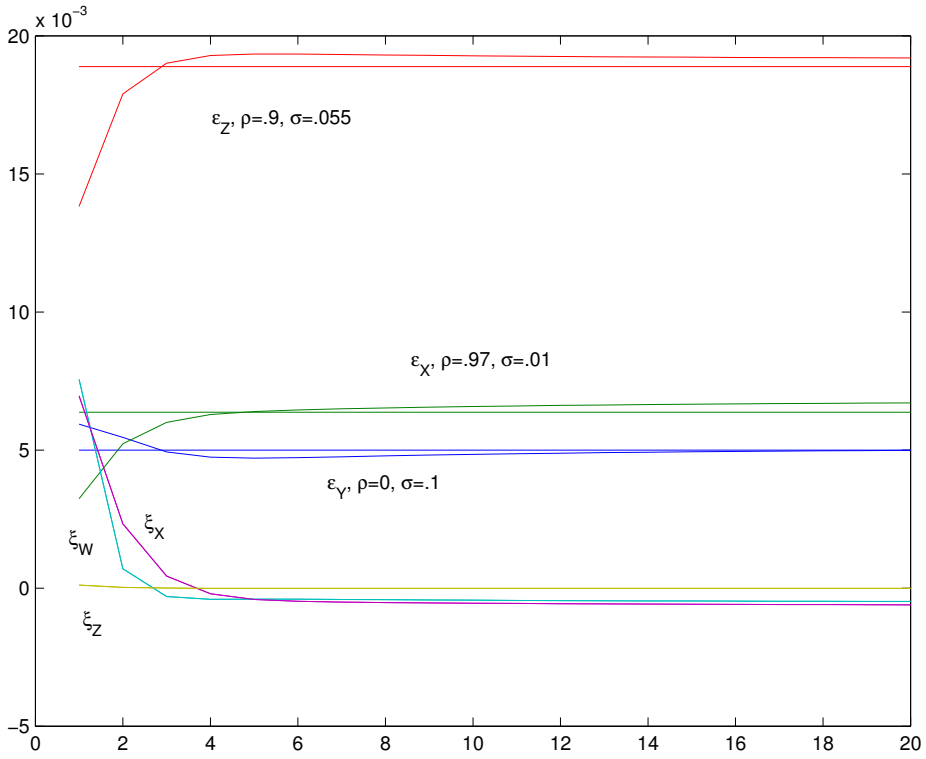

FIGURE 8

Furthermore, if one looks at the diagonal of the matrices of impulse response graphs that emerge from VAR studies, one finds impulse responses that are strongly discontinuous at zero.[12] The adjustment cost models of macroeconomic theory generally imply that variables subject to adjustment costs should respond smoothly to every source of disturbance. To explain the quick estimated responses to own shocks requires introducing disturbances to the adjustment mechanism itself. While it is possible to build a DSGE model containing such disturbances (as shown in the 1998 paper), the distance from microeconomic intuition in the resulting construction is uncomfortably great.

As we have seen in our examples above, information-capacity limits can account simultaneously for smooth response to other variables, real or nominal, and for an idiosyncratically random component of variation, arising from the same source as the smoothness in the response to other variables.

[12]Matrices of impulse response graphs of various sizes can be found, e.g., in Leeper, Sims, and Zha (1996).

The theory does suggest, though, that the randomness in an individual's response to market signals should be truly idiosyncratic, arising from his own internal information-processing constraint. It might be thought, then, that it should vanish, or at least shrink drastically, when aggregated across individuals, so that the responses to own-shocks in VAR's could not come from this source. Recall, though, how the idiosyncratic information-processing randomness arises. It arises from coding inaccuracy, the need to approximate fine-grained information with cruder approximations. People share a need to code macroeconomic data efficiently, and they pay for this service. To a considerable extent, people's needs for coding are similar, so they rely on common sources of coded information. The result is that a considerable part of the erratic response arising from information capacity constraints is common across individuals.

To make this point concrete, consider one of the most pervasive coding services, the daily newspaper, and how it presents macroeconomic data. Many newspapers report the Federal Funds Rate to 3 significant figures every day, at a predictable location in the back of the business section. The vast majority of newspaper readers do not look at this number every day, and of those that do look at the page containing the number, the vast majority make no adjustment in their behavior in reaction to the number. On the other hand, if the New York Times ran as a three-column, front page headline "FED UNEXPECTEDLY RAISES FEDERAL FUNDS RATE 1.5%", many readers of the newspaper would be likely to act on the news. The coding of the time series behavior of the Federal Funds rate into different type sizes at different locations in the newspaper according to its importance is part of the overall information-processing service that a newspaper performs for its readers and that readers (and advertisers) pay for. But, just as with the coding of a Gaussian random variable into a finite set of 0-1 switches introduces a coding error, the newspaper's information processing can influence the erratic component of behavioral response to data. If everyone were tracking and reacting to the Federal Funds rate hour by hour, it would not matter whether the newspaper put it on page one in one inch type, on the front page below the fold, on the first business page, or simply in the usual daily table with no mention in a text story. But in fact the treatment that newspapers (and TV) give this news affects the way people react to it, creating a common component to the idiosyncratic error generated by information-processing. Newspapers are far from the only example of such a phenomenon. Information spreads among people through conversation and imitation, for example, and these channels are themselves subject to coding error that is common across people.

Of course not all of the stochastic component of behavior induced by information-processing is common, and this does imply limitations on the theory as applied to aggregate data. The tight relation between the degree of smoothing in the systematic

component of behavior and the nature and size of the stochastic component that comes out of the examples we have discussed in sections IV and VI will not translate reliably to aggregate data. In particular, we might expect, because of the effects of averaging across agents, that smoothing effects will be stronger relative to the size of idiosyncratic error components than a "representative agent" theory would suggest.

VIII. Comparison to rational expectations theories

Versions of rational expectations that postulate a common information set for all agents at all times imply quick, error-free reactions of all prices and all kinds of agent behavior to every kind of new information and therefore contrast strongly with the implications of rational inattention theory — and with the data. Versions that postulate differing information sets, e.g. between policy-makers and the public or between workers and firms, if they postulate that the difference is in the form of a pure delay in arrival of information, are equally in conflict with rational inattention theory. Delay alone, without smoothing and idiosyncratic error, has no effect on the rate of information flow.

Rational expectations theory based on signal-extraction, like that in Lucas (1973) is much closer to rational inattention theory. Since filtering a behavioral time series Y_t through a one-sided linear filter does not change its information content, in tracking problems like those in section IV the behavior of rationally inattentive agents will always be as if they faced exogenously given noise of the same form as the endogenous noise derived from information-processing constraints. Since rational inattention implies restrictions on the relation between noise and systematic component, we could in principle test for it vs. more general forms of signal-extraction effects, though the problems with aggregation and "indexation" mentioned above imply that this will be difficult.

More importantly, the three types of theory (information-delay RE, signal-extraction RE, and rational inattention) have different implications for how we extrapolate from observed history in projecting the effects of changes in policy that change the dynamic properties of the economy. Information-delay RE suggests that agents act on optimally formed forecasts of the future. It implies that functions relating behavior to optimally-formed forecasts of the future, as generated from the true aggregated model, should remain stable as the dynamic properties of the economy change. Neither signal-extraction RE nor rational inattention theory shares this prediction. Both of these theories postulate the existence of constraints on agents' behavior that makes them act on forecasts that are worse than those that could be constructed from exact knowledge of all the data in the aggregated model.

Signal-extraction rational expectations takes the nature of the noise as exogenously given. It leads, therefore, to the prediction that scaling up the variance of the signal leads to a higher correlation of the signal with agents' behavior, as Lucas (1973) pointed out. Rational inattention theory predicts that agents will behave as if facing noise whose nature changes systematically as the dynamic properties of the economy change. If a fixed amount of capacity is allocated to monitoring the aggregate price level, for example, the amount of noise will rise with the variability of the price level, so that the accuracy of agents' estimates of the inflation rate will not improve, in terms of R^2, as the variance of inflation increases. On the other hand if, as seems more likely, increased variation in inflation is associated with increased marginal returns to estimating it accurately, capacity may be reallocated from other forms of information-monitoring to allow increased accuracy of inflation-monitoring. In this case, the rational-inattention theory provides an explanation for why economic efficiency might deteriorate in the presence of highly variable inflation. It also provides a qualitative explanation for phenomena like the tendency of inflation-indexation clauses in labor contracts to become more prevalent when inflation is variable, but then to disappear again when inflation stabilizes.[13]

IX. Work ahead

To proceed to general equilibrium models incorporating capacity constraints will require attacking some interesting issues. Most important, how do market mechanisms operate in coordinating behavior of finite-capacity agents? There seems to be no technical barrier to working out models of such markets, but there are important questions about how to formulate them. It seems likely that purely competitive exchange markets relating such agents would make actual sales smoother than they would be without capacity constraints and at the same time might lead to more volatile prices than otherwise. But results may depend on whether we think of agent as setting prices in response to quantities or vice versa. With price-setting sellers, the smoothed response of buyers to price changes due to information-processing constraints creates a source of temporary market power. On the other hand, buyers might seek out sellers whose price time series are easier to monitor. Observations like these may eventually lead to more convincing general equilibrium stories about the origins and consequences of price and wage stickiness than we now have, but there is obviously a long way to go.

The models we have considered postulate a limit on information processing capacity, but no limit on agents' abilities to behave optimally given the constraints on capacity. We have split behavior into two levels, an outside-of-time optimization level,

[13]Find reference.

in which optimal rules are derived, conditional on the limitations of real-time data processing but assuming no limits on computational capacity in solving the optimization problem, and the real-time reaction level itself, in which we recognize limits on computational capacity. While this may be a reasonable approximation, it is obviously subject to criticism — there is not in fact any "phase" of our lives during which we optimize behavior without constraints of time or computational capacity. There is previous work in the game theory literature (Abreu and Rubinstein, 1988) that works with the notion of agents as "finite automata". This approach also splits the people it models into an unboundedly rational optimizing level and a computationally constrained level, but it postulates a very different dividing line between levels.

Of course this is only the tip of the iceberg of ways this theory is still incomplete.

X. Conclusion

It seems presumptuous to have a section titled "Conclusion" in a paper like this that consists mostly of thinly supported speculation. This paper has improved on the even vaguer speculations in Sims (1998) by showing that a capacity constraint can substitute for adjustment costs in a dynamic optimization problem. In several respects the resulting setup seems more realistic than adjustment cost frameworks, and there are interesting differences in its implications. It is probably worthwhile to work on this a little further.

Appendix A. Solution methods

References

Abreu, D., and A. Rubinstein (1988): "The Structure of Nash Equilibrium in Repeated Games with Finite Automata," *Econometrica*, 56(6), 1259–1281.

Bénabou, R., and J. Tirole (2001): "Self-Knowledge and Self-Regulation: An Economic Approach," Discussion paper, Princeton University.

Evans, G. W., and S. Honkapohja (2001): *Learning and Expectations in Macroeconomics*. Princeton University Press, Princeton, NJ.

Giannoni, M. P. (1999): "Does Model Uncertainty Justify Caution? Robust Optimal Monetary Policy in a Forward-Looking Model," discussion paper, Princeton University.

Gray, R. M., and D. L. Neuhoff (2000): "Quantization," in *Information Theory: 50 Years of Discovery*, ed. by S. Verd/'u, pp. 281–339. IEEE Press, Piscataway, NJ, All articles reprinted from *IEEE Transactions on Information Theory* 4, October 1998.

Gul, F., and W. Pesendorfer (2001): "An Economic Theory of Self Control," *Econometrica*.

Hansen, L. P., and T. J. Sargent (2001): "Robust Control and Model Uncertainty," *American Economic Review*, 91(2), 60–66.

Keating, J. W. (1997): "Is Sticky Price Adjustment Important for Output Fluctuations?," Discussion paper, University of Kansas, University of Kansas.

Laibson, D. (1997): "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics*, 112, 443–478.

Leeper, E. M., C. A. Sims, and T. Zha (1996): "What Does Monetary Policy Do?," *Brookings Papers on Economic Activity*, (2).

Lucas, Robert E., J. (1973): "Some International Evidence on Output-Inflation Tradeoffs," *American Economic Review*, 63(3), 326–334.

Onatski, A., and J. H. Stock (1999): "Robust Monetary Policy Under Model Uncertainty in a Small Model of the U.S. Economy," manuscript, Harvard University.

Sargent, T. J. (1993): *Bounded Rationality in Economics*. Oxford University Press, Oxford.

Schervish, M. J. (1995): *Theory of Statistics*, Springer Series in Statistics. Springer, New York.

Sims, C. A. (1998): "Stickiness," *Carnegie-rochester Conference Series On Public Policy*, 49(1), 317–356.

Department of Economics, Princeton University
E-mail address: sims@princeton.edu