

RATIONAL INATTENTION

1. MOTIVATION

Introspection:

- We know we don't respond to all information available to us without significant economic cost.
- The price you are willing to pay for a sandwich at lunch should depend — at least a tiny bit — on what has happened to the short term interest rate this morning. But almost no one makes this connection.

2. MOTIVATION

Anecdotal Observation:

- This might be in part because because the price posted for each sandwich usually is the same from day to day, and when it does change, changes in a discrete jump.
- But these prices should be varying — at least a tiny bit — with changes in the interest rate, the futures price of beef, wheat and corn, and lots of other things.
- Is this because customers have a hard time dealing with changing prices? Because they would not react to small changes? Because the restaurant owner does not find it worthwhile to keep track of and respond to every source of changes in cost at every moment?

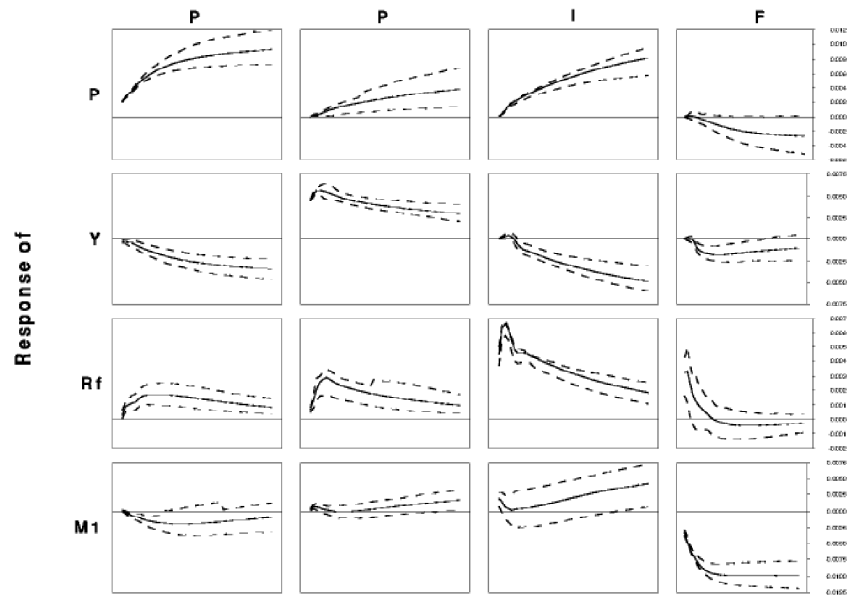
3. MOTIVATION

Macroeconomics:

- Theorists complain that the Smets-Wouters model, and its descendants and ancestor (Christiano, Eichenbaum and Evans), are too full of “ad hoc” sources of inertia.
- These sources of inertia are there because without them, economic theory implies much faster reactions of economic variables to disturbances than we see in the data.
- Keynes just assumed prices move slowly. Neo-Keynesians “micro-found” Keynes by assuming there are costs of adjustment of prices, or “technological” barriers to frequent changes in them. But there is little empirical microfoundation for many of these costs.

4. MOTIVATION

Figure 5
Four Variable Restricted Model



5. THE BASIC IDEA OF RATIONAL INATTENTION

- A person responding to randomly fluctuating market signals and translating them into actions has some of the characteristics of an engineers communications channel: There are inputs (market signals), outputs (actions) and a maximum rate at which inputs can be translated into output without significant error.
- An internet connection has a **capacity** measured in bits per second (or, nowadays, more likely megabytes per second (MB/s or MiB/s)). Surely a person translating market signals to actions faces a similar, though probably tighter, limit.

6. CONTRAST WITH VALUE OF INFORMATION IN STANDARD DECISION THEORY

Standard setup:

$$\max_{\delta} E[U(Y, X, Z, \delta)] \quad \text{subject to}$$

$$\text{either } \begin{cases} \delta \text{ a function of } X, Z \\ \delta \text{ a function of } X \end{cases} .$$

Difference in $E[U]$ between the two cases is the “value of the information in Z ”.

7. ENTROPY, DISCRETE CASE

- **Entropy** is a measure of “how much uncertainty” there is in a probability distribution.

- It depends on the *distribution*, or probability measure, not on the values taken on by a random variable.
- If heads or tails each has probability .5, and next year's profits have probability .5 of being \$1 million or \$1.75 million, both random variables have distributions with the same entropy.
- It is $-\sum p_i \log(p_i) = -E[\log(p_i)]$, where i enumerates the points in the probability space.

8. ENTROPY, DISCRETE CASE

Axiomatics: We assume first that we want the “amount of information” obtained by resolving the uncertainty in a distribution over a finite state space to depend only on the probabilities $\{p_i, i = 1, \dots, n\}$ of the points in the space. Call this function $H(p_1, \dots, p_n)$

Then we add the requirement that the uncertainty is additive in the following sense. Suppose $\{X_1, \dots, X_m\}$ is a sequence of random variables defined on the n -point finite state space with probabilities defined by $\{p_1, \dots, p_n\}$. Assume $Y = f(X_1, \dots, X_m)$ is a function of these random variables with the property that the Y random variable has a unique value at each of the n points in the state space — meaning that observing Y , or the values of $\{X_1, \dots, X_m\}$ is enough to reveal which point in the state space has been realized. After seeing X_1 , our distribution over the state space is $p(i | X_1)$. After seeing X_1 and X_2 it is $p(i | X_1, X_2)$, etc. Since observing Y is equivalent to observing the value of the state, we will use the notation $H(Y)$ as equivalent to $H(p_1, \dots, p_n)$ and $H(Y | x)$ for the entropy of the distribution of Y conditional on a particular value x of the random variable X . $E_X[H(Y | x)]$, where the expectation is over the values of the random variable X , we write as $H(Y | X)$. We would like to have this property:

$$H(X_1) + H(X_2 | X_1) + \dots + H(X_n | X_1, \dots, X_{n-1}) = H(Y)$$

This property is enough to tell us that $H(Y) = H(p_1, \dots, p_n) = -E[\log(p_i)]$. We can choose the base of the log, but otherwise the measure is unique. If we use log base 2, the unit of measurement is called a “bit”. One bit is the information in a flip of a fair coin. If we use natural logs the unit is called a “nat”. Proving this result is not too hard, if one adds the assumptions that H is continuous in \vec{p} and that adding points to the probability space that have probability zero does not change the entropy.

9. ENTROPY, GENERAL CASE

- Entropy can be defined for distributions on any probability space.
- It is always defined relative to some base measure. For finite probability spaces, the base measure is counting measure and the density is the usual $\{p_i\}$.
- For continuously distributed random variables, The standard base measure is Lebesgue measure on \mathbb{R}^n , so entropy becomes $\int_{\mathbb{R}^n} p(x) \log(p(x)) dx$.

10. MUTUAL INFORMATION

- If X and Y are two random variables with a joint distribution defined by a density $p(x, y)$ over a base measure $\mu_x \times \mu_y$, then the **mutual information** between X and Y is $I(X, Y) = H(X) + H(Y) - H(X, Y)$.

- Equivalently, it is $H(Y) - H(Y | X)$ or $H(X) - H(X | Y)$. That is, it can be thought of as the expected reduction in the entropy of Y from observing X , or equivalently as the expected reduction in the entropy of X from observing Y .
- Note that $I(X, Y) = H(X)$ if and only if $H(X | Y) = 0$, i.e. if and only if there is no uncertainty about X after we observe Y .

11. CODING

Suppose we have a record of 1024 (2^{10}) time periods in which 5 events occurred. Our raw data is a string of 1024 0's and 1's, with the 1's in the positions corresponding to the dates of the events. We want to send this via a telegraph key that each quarter second can reliably transmit a 0 or a 1.

If we just send the raw data, it will take $1024/4 = 256$ seconds, or about 4 minutes. But of course we could instead agree with the person receiving the information that we are not sending the raw data, but just the positions in the sequence of the ones. Each such position can be represented as a number between 1 and 1024, and in binary notation these are 10-digit numbers. Since there are just five one's, we can get the message through with a transmission of 50 0's and 1's, or about 12 seconds.

We are **coding** the raw data into a different form that can be more efficiently transmitted through our telegraph key.

This coding scheme is a good one if the 1's are rare and randomly scattered through the sequence. If the 1's formed more than about 10% of the raw data, transmitting all their locations this way would take longer than transmitting the raw data. So our ability to code it to achieve faster transmission depends on the probability distribution from which the raw data is drawn.

If the message is always going to be five event dates, randomly drawn from the integers 1 to 1024, then the entropy of the distribution of messages is five times the entropy of a distribution that puts equal probability on 1024 points, i.e. $5 \log_2(1024) = 50$ bits. Or if the events are i.i.d. across dates, with probability $5/1024$ of a one at each date, the entropy of each draw is .0445 and thus the entropy of the whole sequence is $1024 \times .0445 = 45.64$ bits.