1. A Digression on Information Theory

In this section it is argued that empirical observations of delayed, smooth reactions of variables in economic models to one another, combined with non-smooth responses of these variables to "own shocks", fits the hypothesis that individuals have limited information processing capacity. Under this hypothesis, the path connecting market signals to individuals' behavioral reactions should have the characteristics of a finite capacity *channel*, in the language of information theory.

A channel is something that takes as input a *signal* and produces an output. Once we have defined the possible input signals and described how they map into outputs (including the possibility that the mapping involves random noise) we can calculate what is called the channel's *capacity.* The central result of information theory is that it is possible to characterize the amount of information in any message we might want to transmit, and that if we allow a time for the transmission process equal to (amount of information)/(capacity), we can send the message with arbitrarily small error, even if the channel itself has substantial random noise in the connection of signal to output.

To make this concrete, we start with the simplest possible case. First we need to define the amount of information in a message. A message has to be characterized as the realization of a random variable. Or, in other words, we have to define the range of possible messages and their relative probabilities in order to define the amount of information in the message. In this simplest case, the message may be either 1 or 0, with probability $P[1] = p$. We characterize information by first characterizing the amount of lack of information, or *entropy* in a probability distribution. It turns out to be convenient to measure the entropy of a random variable z as $-E[\log_2(p(z))]$, where p is the p.d.f. of z. The use of base 2 logs is arbitrary, and the base measure against which the p.d.f. is constructed is also arbitrary, but otherwise it is possible to show axiomatically that this is the only reasonable measure of entropy.[1] Thus for our

[1] If we change the base measure, the ranking of distributions by entropy will change. For example, our base measure, instead of putting equal discrete weight on 0 and 1, with no measure elsewhere, (as is implicitly assumed when we conclude that with $p = .5$ the distribution has 1 bit of entropy), could put weight 2 on x=1 and weight 1 on x=0. Then the p.d.f. of a distribution with $P[1] = .5$ would take the value .25 at $x = 1$ and .5 at $x = 0$ and the entropy of the distribution would be $I(p) + p$, where $I(p)$ is the standard definition of entropy which is 1 at $p = .5$. With this definition of entropy maximum entropy occurs not at $p = .5$, but at $p = \frac{2}{3}$. Nonetheless, it can be verified that the average rate of information flow through a channel that transmits 0 or 1 without error, drawing from a distribution with $p = .5$, is still 1 bit per second with this modified definition of entropy. More generally the base measure has no effect on the average rate of information flow, no matter what the distribution of s and no matter what the distribution of $x|s$, and thus no effect on the definition of channel capacity.

two-point distribution, the entropy is $p\log_2(p) + (1-p)\log_2(1-p)$. This expression converges to zero as p approaches zero or one, and is maximized at $p = .5$, where it is just equal to one. The entropy of a two-point distribution with equal probabilities on the two points is called one *bit* of information.

Now we consider the simplest possible channel. Time t is discrete. The channel can take as its signal s at any date t only two possibilities: $s = 1$ or $s = 0$. The mapping from signal to output x is trivial: $s = 0 \Rightarrow x = 0$, $s = 1 \Rightarrow x = 1$. In other words there is no random error. If at each date we send a signal s drawn from a two-point distribution with equal probabilities on 0 and 1, then the receiver of the output each period has his distribution over s converted from one with entropy 1 bit (before he sees the signal) to one with zero entropy (after he sees the signal). Thus the channel transmits one bit per time period.

Of course we could also draw s from a distribution with $p \neq .5$, in which case less than one bit per time period would be transmitted. This suggests that we are wasting channel capacity in transmitting this way. But suppose we actually need to send a stream of zeros and ones in which, say, 90% of the stream is ones. How can we avoid wasting capacity? Break the stream of zeros and ones into groups of, say 4. Map these sequences into sequences of s for transmission as follows:

message	s sequence	n	p	$n \cdot p$
1111	1	1	0.66	0.66
0111	0000	4	0.073	0.292
1011	0001	4	0.073	0.292
1101	0010	4	0.073	0.292
1110	0011	4	0.073	0.292
0011	01000	5	0.0081	0.0405
0101	01001	5	0.0081	0.0405
0110	01010	5	0.0081	0.0405
1001	01011	5	0.0081	0.0405
1010	01100	5	0.0081	0.0405
1100	01101	5	0.0081	0.0405
1000	011100	6	0.0009	0.0054
0100	011101	6	0.0009	0.0054
0010	0111100	7	0.0009	0.0063
0001	0111101	7	0.0009	0.0063
0000	0111110	7	0.0001	0.0007
mean no. of s's per 4 message elements		2.0951		

It is easy to see that this scheme transmits the message without error, while at the same time requiring about half as much time to send a given message, on average, as the naive method of sending an s sequence that matches that in the message. The naive system would send $-.9\log_2(.9) - .1\log_2(.1) = .47$ bits per time period, whereas the scheme in the table sends almost twice that, or about .9 bits per time period. This is not quite full channel capacity, of course. To get closer to full capacity we would have to create a coding scheme for message sequences longer than 4.

Suppose the channel contains noise. For example, suppose sending $s = 1$ results in output $x = 1$ only with probability .75, while with probability .25 it results in $x = 0$. Suppose that the probability of "error" is the same when a 0 is sent. Then a channel user whose distribution on s was equal-probability on zero and one before seeing x, would after seeing x have a conditional probability distribution characterized by $P[s = x|x] = \frac{2}{3}$. The entropy of this distribution is .9183, so the information transmitted is 1 (the entropy of the distribution before seeing x) minus .9183, or .0817. This channel transmits .0817 bits per unit time. This means that if we are willing to transmit about 1/.0817, or about 13, s's for every 0 or 1 in a message sequence that is half ones and half zeros, we can transmit the message at that rate with arbitrarily small error. The type of code that does this is a bit more complicated than that shown above, so no example is provided here. Texts in information theory or computer science discuss such “error-correcting” codes.

In macroeconomic modeling, the messages agents send each other via market signals are generally most naturally thought of as real numbers, not zeros and ones. If s is drawn from a probability distribution with a density on the real line, and if the channel is the kind of simple error-free channel we discussed first above, with $s = x$ with probability one, the channel is physically unrealizable, because it has infinite capacity. This may seem paradoxical, but a real number, transmitted exactly, amounts to transmitting an infinite sequence of 0's and 1's without error. Every infinite sequence of 0's and 1's can be thought of as the binary representation of a real number between 0 and 1, so we can send any message of 0's and 1's through this channel in one time period, no matter how long the message.

But if the channel contains noise, it can transmit real-valued s, producing real-valued x, while transmitting a finite amount of information per unit time. For example, suppose $s = x + \varepsilon$, where ε is i.i.d. $N(0,1)$ and our distribution over s before seeing x is $N(0,1)$. The distribution of s conditional on x is $N(.5s,.5)$. The entropy of a $N(0,\sigma^2)$ distribution is

$$\frac{\log_2 e}{2} + \log_2 \sigma + \frac{\log_2 \pi + 1}{2} \,.$$

Therefore the information transmitted by observation of x is $\log_2 \sqrt{1} - \log_2 \sqrt{.5} = .5$ bits.

Note that it cannot be that the channel is characterized by the ability to transmit an s drawn from an arbitrary distribution on the real line with an additive error, since if this were true the relative error could be made arbitrarily small by coding so that the s's actually sent are drawn from distributions that make s with high probability either very large or very small. The channel would have to require either a bound on the range of s, or an error specification that makes the distribution of the error depend on the distribution of s.[2] But our interest here is in characterizing the information transmission rate implicit in a given relation of inputs signals to outputs, not in an analysis of optimal coding for given physical channel characteristics. With proper coding, transmission of a $N(0,1)$ signal with an additive $N(0,1)$ error can be

[2] The error must be specified so that it remains possible to confuse two s values, no matter how far apart they are. For example making the standard deviation of ε proportional to $1+s$ works.

accomplished with a channel that sends one 0 or 1, without error, per unit time, and the transmission can be at a rate of two signals per unit time. This is just another way of stating the fact that the information transmission is at the rate of .5 bit per time period.

We conclude that finite-capacity transmission of real numbers must involve random error. Now we proceed to argue that transmission of information in continuous time must also in a certain sense involve delay.

Suppose we have a signal s that follows a standard Wiener process, and that the output of information transmission is $x = s + \varepsilon$, with ε also following a standard Wiener process, independent of s. This implies transmission of information at an infinite rate. To see why, consider the changes in s, x, and ε over small time intervals $((j-1)/n, j/n$. We use the notation that, e.g., $\Delta s_j = s(j/n) - s((j-1)/n)$. The changes Δs_j and $\Delta \varepsilon_j$ will be i.i.d. across time, at any one date t being uncorrelated with each other, jointly normal, both distributed as $N(0, 1/n$. The change in x over the interval will therefore be distributed as $N(0, 2/n)$. But then if Δs_j is treated as the signal and Δx_j as the output, we are in the situation already discussed, where observation of the Gaussian output reduces the standard deviation of the distribution of the unknown Gaussian signal by a factor of 2, and a half bit of information is transmitted. This half bit per transmission is the same no matter how large n has been chosen. Thus in any finite interval, arbitrarily large amounts of information can be passed through this channel. This will be true in any setting where the error and the signal are independent stochastic processes driven by stochastic differential equations of the same order with Wiener process forcing terms.

A situation where, in continuous time, information is transmitted at a finite rate, can be constructed as follows. Let s be a cumulated Wiener process, i.e.

$$s(t) = \int_0^t W(v)\, dv \;, \tag{1}$$

and let ε be a Wiener process independent of s. Consider what happens when, as above, we break a finite interval, say (0,1), into n equal subintervals and examine the joint distribution of the sequences of s_j's and x_j's, where (for example) $s_j = s(j/n) - s((j-1)/n)$. We have to use the full joint distribution of the sequences, rather than examine the information transmitted at each j, one at time, because here the transmission error and the signal have different serial correlation properties, so no differencing or other filtering operation can produce i.i.d. signal-output pairs. The information about the sequence $\{s_j\}$ transmitted by observation of the sequence $\{x_j\}$ can be found by computing the entropy of the unconditional distribution of the full s sequence and comparing it to the entropy of the conditional distribution, given the x sequence. Both distributions will be multivariate normal, and it turns out that the information transmitted is

$$\tfrac{1}{2}\log_2\left(|\Sigma_s| - \left|\Sigma_{s|x}\right|\right)\ ,$$

where Σ_s is the covariance matrix of the unconditional distribution of the s_j sequence and $\Sigma_{s|x}$ is its covariance matrix conditional on the x_j sequence. This can be calculated directly, and it can be shown to converge to a constant as $n \to \infty$. The convergence is rapid. Simply treating the signal as the single value $s(1)$ and the output as $x(1)$ transmits .415 bits, breaking the interval into two pieces ($n = 2$) transmits .613 bits, $n = 20$ transmits .721 bits and $n = 80$ transmits .727 bits. Evidently, with this smooth signal and less smooth error, there are rapidly diminishing returns to sampling x more frequently.

These two examples are representative. When s and ε are generated by nicely behaved differential equations driven by Wiener processes, then if ε has at least as many derivatives as s, observation of x can transmit unbounded amounts of information in finite time, while if ε has fewer derivatives than s, observation of x over a finite interval can transmit only finite amounts of information, no matter how frequently x is sampled.

Suppose there is a signal, say an asset price Q, that we wish to model as influencing behavior via a finite-capacity channel, with the behavior being represented by a choice variable Y. The implication of the analysis we have just been through is that we expect the relation between Y and Q to take the form (if we are modeling with Ito processes)

$$dY = a(Q)dt + dw(t) \tag{2}$$

with the W in (2) independent of that driving Q, or (if we are working with general linear systems)

$$Y(t) = \int_0^\infty a(s)Q(t-s)\,ds + \varepsilon(t) = a * Q(t) + \varepsilon(t)\ , \tag{3}$$

where the $*$ denotes convolution and ε is a white noise independent of Q. If instead the variance of ε or W were zero, or if the level rather than the time derivative of Y depended on Q, the relationship would imply information flow at an infinite rate.