WHY RATIONAL INATTENTION IMPLIES SMOOTHED AND DELAYED RESPONSES

Suppose Y_t is a given stochastic process and X_t is the stochastic process of our decisions made based on observations of Y that are subject to a Shannon capacity constraint. To keep things simple initially, suppose Y is a Wiener process with innovation process the white noise ε_t. This can be written using two different notations:

$$Y_t = \int_0^t \varepsilon_{t-s}\, ds \text{ or} \tag{1}$$

$$dY_t = \varepsilon_t\,. \tag{2}$$

We also suppose for simplicity that our objective is to minimize $E[(Y_t - X_t)^2]$.

In a static version of this problem with an information constraint and Gaussian Y, where we are observing a single value Y with finite capacity, we get a solution in which $Y = X + \xi$, where ξ is normal with zero mean and independent of X. We can think of this solution as generated by observing $Z = Y + \nu$, where ν is independent of Y, and then setting X to be a linear function of Z.

In this dynamic problem, you might then initially expect the solution to be interpretable as observing $Z_t = Y_t + \nu_t$, with ν_t independent of $\nu_t$and itself a Wiener process. Indeed if one could make such observations, the optimal solution would have the form $Y_t = X_t + \xi_t$. However, making such observations would imply an infinite rate of information flow between the processes for X (or Z) and Y.

The entropy of a Gaussian random vector A with covariance matri Σ is

$$-\int_{-\infty}^{\infty} \phi(a;\mu,\Sigma)\big(-\tfrac{1}{2}(a'\Sigma^{-1}a + \log(2\pi) + \log|\Sigma|\big)\, da \tag{3}$$

where $\phi(\ ;\mu,\Sigma)$ is the normal density with mean μ and variance matrix Σ. This is easily verified to be a constant plus $\frac{1}{2}\log|\Sigma|$. (This is entropy in nats. To get bits, replace log with $\log_2$.) And of course if A is one-dimensional this is just a constant plus the log of the standard deviation of A.

Thus if the pair of random variables A, B are jointly normal, the mutual information between them is

$$H(A) - H(A \mid B) = \log\sigma_A - \log(\sigma_A \mid B) = -\tfrac{1}{2}\log(\rho^2)\,, \tag{4}$$

where ρ is the correlation between A and B. Note that this means rescaling the variables does not change their mutual information.

Now consider our hypothetical solution with $Y_t = X_t + \xi_t$, ξ_t independent of X_t. Consider the pair $Y_{t+\delta} - Y_t$, $X_{t+\delta} - X_t$. Because X, Y and ξ are all by construction Wiener processes, they have stationary increments. Furthermore, it is easy to see that the correlation of $Y_{t+\delta} - Y_t$, with $X_{t+\delta} - X_t$ does not depend on the interval δ over which we take differences. This means the mutual information between these differences is the same, whatever δ we consider. Thus by choosing δ small enough, we can transmit arbitrarily many bits in any given finite time interval. That is, the rate of information flow between X and Y is infinite.

If we describe the joint behavior of X and Y with a first-order linear stochastic differential equation of the form

$$d\begin{bmatrix} Y \\ X \end{bmatrix} = \alpha \begin{bmatrix} Y \\ X \end{bmatrix} dt + \beta dZ\,, \tag{5}$$

where Z is a two-dimensional Wiener process and α and β are 2×2 matrices of coefficients, the same kind of argument we gave to claim $X = Y + \xi$ implied an infinite rate of information flow implies that the β matrix must be diagonal if the Y, X pair in (5) is to imply a finite rate of information flow. That is, if we are trying to keep X close to Y, the only route open to create any correlation between the two processes, with finite information flow, is through the α matrix.

Using alternative notation, we can write the joint Y, X process as

$$\begin{bmatrix} Y_t \\ X_t \end{bmatrix} = \int_0^\infty A(s) \begin{bmatrix} \varepsilon_{t-s} \\ \xi_{t-s} \end{bmatrix} ds = A * \begin{bmatrix} \varepsilon \\ \xi \end{bmatrix}(t)\,, \tag{6}$$

where ε and ξ are "fundamental" driving white noise processes for Y and X. (What "fundamental" means here is too complicated to explain in this course, if you have not seen an explanation before. Any Gaussian stationary process, or even any process with stationary increments, can be given a representation in this form.) In this notation, finite information flow requires that, if A_0 is non-singular, it is diagonal. More generally, if the elements of the $A(s)$ matrix-valued function have finite right derivatives at zero, finite information flow rate requires that the lowest-order right derivative of A at zero that is non-zero must be diagonal.

More intuitively, this says that for finite mutual information flow, it must be that X is the sum of one piece that is correlated with Y but has one more time-derivative than Y, and another piece, orthogonal to the first, that has the same number of derivatives.

Finally, one more way to state this point is that the array of impulse responses in a bivariate VAR for Y, X must be smoother and more delayed off diagonal than on the diagonal.