SEVEN ESSAYS ON MINIMUM ENTROPY

Jon F. Claerbout

A geophysicist peering into a microscope viewing biological tissue will have no trouble focusing the microscope. The focusing is not done by measuring the focal length of the lenses and matching the distance to eye and specimen. The focusing is done by enhancement of some characteristic of the image. This is possible despite the likelihood that the geophysicist has little or no previous experience with the image or microscope. What characteristic of the image is sought? Perhaps it is short spatial wavelengths, perhaps bandwidth in spatial spectra, perhaps dynamic range in intensity. In Minimum Entropy (ME) data analysis research we try to determine physical parameters (such as distance or focal length) by means of adjustments which sharpen an image. Despite visions of scientific precision conjured up by the word "entropy" our work is still largely empirical, though I believe it is very promising. These essays give an account of current efforts to apply and systematize this kind of inductive learning in reflection seismology. The essays are largely independent of one another. They may be read out of sequence and without reference to earlier work. Titles with brief descriptions are:

1. Applications of ME Processing - A new family of processes in reflection data analysis are possible.

2. Geometric Inequality versus Power Inequality - A comparative analysis of two previous approaches.

3. The Basic Debubble Algorithm - How to do it.

4. Non-stationarity: Application to a Common-Shot Profile - An attempt to specify a production program.

5. Seismogram Inversion - How to use ME to zap multiple reflections.

6. ME Extrapolation and Spectrum - Counterpoint to Burg's *maximum* entropy method.

7. Convex Inequalities and Statistical Mechanics - Physicists and chemists had been using the concept of entropy long before information theorists took it up. Here is what they mean by it and how it relates to our imaging concepts.

1. Applications of ME Processing

The major ME success to date has been debubbling. Is it a fortunate accident that the best subjective view of biological tissue gives an accurate matching of the focal length to distance? The results of analysis of synthetic data and the results of near-field recording show a similar fortunate accident, that source waveforms found by ME in the far field usually provide an accurate match to the correct waveforms. Errors are very much smaller than those of the older least-squares methods. Indeed, judged by the new standards of ME, it is fair to say that "least squares doesn't work." The successful application of ME to debubbling points out the potential for many other applications of ME ideas in geophysical data analysis.

Time-variable deconvolution. Absorption is a convenient parameter to vary, and it is simple enough to test the entropy of several different absorptions and select the best. Cosmetically the effect would be that of time-variable filtering. Assuming the validity of that time-honored process, the results should be improved whether or not the "best" absorption turns out to be related to dissipation or to some other poorly understood scattering phenomenon.

Coloring in k_y***-space.*** The square-root expansion used in migration contains ω, hence possible energy absorption (or growth) in each term. The main

absorption occurs in the so-called 5-degree term, as above. A different role is played by the ω in the diffraction term (see Morley, SEP-16, p. 109). Adjustment of this term can have the cosmetic effect of recoloring in k_y-space. (Actually it would be k_y^2/ω-space.) Why not give it a try? The best value, once determined, can be physically interpreted as compensation for the differential absorption between hyperbola tops and flanks. Anyway, even if the physics is inappropriately understood, maybe it will boost up some of those fault-plane reflections which have been suppressed by CDP stack.

Velocity. Iterative adjustment of the migration velocity may be very economically achieved by a number of methods. Cosmetically this is a phase adjustment in k_y-space.

Inversion. Iterative adjustment of reflection coefficients so as to predict and hence subtract multiples is the most ambitious of the proposed applications (because there are the most adjustable parameters). But the number of adjustable parameters is not really very many more than are found in the debubbling process, which is already very successful.

All of the above suggestions amount to saying, "take physical parameters to vary, then use minimum entropy to determine values." Some improvements may be barely perceptible, best seen with "unclipped" displays. Sometimes the processing sequence may be important. All are worth trying, and others will be found.

2. *Geometric Inequality versus Power Inequality*

In SEP-13 an ME deconvolution program was developed based on the power inequality

$$\left[\frac{1}{N}\sum_{i=1}^{N} p_i^{1+\epsilon}\right]^{1/(1+\epsilon)} \geq \frac{1}{N}\sum_{i=1}^{N} p_i \qquad (2.1)$$

Let ϵ approach zero and constrain the p_i to sum to unity. Then, forcing this inequality towards equality turns out to amount to extremalizing $\Sigma\, p \ln p$. It resembles entropy.

In SEP-15 another ME deconvolution program was developed based on the *geometric inequality*

$$\prod_{i=1}^{N} p_i^{1/N} \leq \frac{1}{N} \sum_{i=1}^{N} p_i \tag{2.2}$$

Constrain the p_i to sum to N. Then, forcing this inequality towards equality amounts to extremalizing $\Sigma \ln p$. Defining the p_i as seismogram power as a function of time t=i we could properly describe this process as extremalizing information defined by counting bits. Alternately, sorting the seismogram from smallest to largest and defining the positive intervals as p_i, the sum of $\ln p_i$ could be regarded as an estimate of the expectation of log probability - in other words, Shannon's definition of expected information $E[\ln(p)] = \int p(x) \ln[p(x)]\, dx$. Satisfactory deconvolution programs were based on either (2.1) or (2.2). Although I tried very hard, I was unable to develop a satisfactory program based on the Shannon approach. No one at SEP has yet tried inserting estimated probability functions into (2.1). At this point the geometric inequality seems to have a better philosophical basis than the parsimony or power inequality. But the geometric inequality does have a flaw. Since we are interested in minimum entropy, that is, driving the inequalities as far away from equality as possible, you can see that a single vanishing p_i on the left side of (2.2) could cause trouble. Trouble may be reduced by having p_i refer to a cell of sufficient size, or by incorporating some kind of noise threshold, but these raise more questions than they solve. That is why it is worthwhile to re-examine the philosophical basis for the power inequality (2.1) as ϵ goes to zero. Repeating the derivation using a bit more care for clarity and the scale factor, we take the log of (2.1)

$$\frac{1}{1+\epsilon} \ln \frac{1}{N} \sum_{i=1}^{N} p_i^{1+\epsilon} \geq \ln \frac{1}{N} \Sigma p_i \tag{2.3}$$

Take a Taylor series on the left side about $\epsilon=0$ and subtract the lead term on both sides.

$$\epsilon \frac{\partial}{\partial \epsilon} \frac{1}{1+\epsilon} \ln \frac{1}{N} \Sigma p_i^{1+\epsilon} \geq 0 \tag{2.4}$$

Note that for any constant a

$$\frac{d}{dx} a^u = \frac{d}{dx}\left(e^{\ln a}\right)^u$$

$$= \ln(a) \left(e^{\ln a}\right)^u \frac{du}{dx}$$

$$= \ln(a)\, a^u \frac{du}{dx} \qquad (2.5)$$

Divide (2.4) by ϵ. Differentiate using (2.5). Then let $\epsilon \to 0$ to get

$$- \ln \frac{1}{N} \Sigma p_i + \frac{\Sigma p_i \ln p_i}{\Sigma p_i} \geq 0 \qquad (2.6)$$

Scaling through by Σp we now define the "negentropy" S as

$$S = \sum_{i=1}^{N} p_i \ln p_i - \left(\Sigma p_i\right) \ln\left(\frac{1}{N} \Sigma p_i\right) \qquad (2.7)$$

As a quick check, if all the p_i are identical, $S = 0$.

An important property of this entropy that was not mentioned in SEP-13 is that it is ***extrinsic***. Specifically, let p_i be the square of a seismogram at time $t=i$. Let seismogram number 1 have length N_1, energy $u_1 = \sum_i p_i$, and $s_1 = \sum_i p_i \ln p_i$. Likewise let seismogram number 2 have N_2, u_2, and s_2. By equation (2.7) the separate and combined seismogram entropies are

$$S_1 = s_1 - u_1 \ln\left(\frac{u_1}{N_1}\right)$$

$$S_2 = s_2 - u_2 \ln\left(\frac{u_2}{N_2}\right)$$

$$S_{(1+2)} = s_1 + s_2 - (u_1 + u_2) \ln\left(\frac{u_1 + u_2}{N_1 + N_2}\right)$$

Now it is easy to show that if the energy per unit length on the first seismogram u_1/N_1 equals that on the second seismogram u_2/N_2, then

$$S_1 + S_2 = S_{(1+2)}$$

This means that entropy does not change if a homogeneous region is divided into hypothetical bins. It would change only if the bins got so small that the energy per unit length began to fluctuate.

An important practical factor influencing one's choice of entropy measure is the behavior of derivatives. Stable derivatives are very helpful in finding descent methods that actually work. The geometric inequality has the unfortunate property that $\partial S/\partial p_t$ contains $1/p_t$. Thus a single cell containing no energy has a drastic effect. If this can happen because of random fluctuation, then it is hard to descend. The power (or parsimony) inequality has well-behaved first derivatives, but $1/p_t$ occurs in the second derivative. For that reason the algorithm in the next essay does not use a second-derivative method to determine the step length. [Incidentally, in SEP-20, p. 226-227, there should be "double dot" over p in equations (6c), (7c), and twice in (8c), but since second-derivative methods are unreliable, we will not issue an errata but instead look for the bug in our typesetter!]

In conclusion, the "parsimony" approach has by no means been superseded by the geometric inequality approach. In fact, I should have made reference to it in the SEP-20 paper on separation of binary mixtures. For the future, these two approaches seem quite different and much remains to be learned about their properties. Most tantalizing of all is the still unreached goal of successful use of the probability entropy in seismogram analysis.

3. The Basic Debubble Algorithm

The basic algorithm for determination of the minimum entropy deconvolution filter is an interesting combination of well-known techniques for least-squares optimal filters. Define

y_t = observed seismogram

f_t = decon filter = inverse bubble

x_t = $y * f$ = output deconvolved trace

Y = matrix of shifted columns of y_t such that $x = Yf$

p_t = x_t^2 or smoothed version of same or envelope

g_t = weights

G = diagonal matrix with g_t on diagonal

As a classical least-squares situation, consider a weighted sum of output powers. It has summation representation and matrix representation.

$$S' = \sum_t g_t p_t = x^T Gx = f^T Y^T GYf \tag{3.1}$$

To extremalize S' with respect to variation in the filter, we set to zero

$$\frac{\partial S'}{\partial f_j} = \sum_t g_t \frac{\partial p_t}{\partial f_j} = 2Y^T GYf = 2Y^T Gx \tag{3.2}$$

As a second example consider some "entropy function" $S(p)$ such as the extrinsic power inequality form

$$S = \sum_{t=1}^{N} p_t \ln p_t - \left(\sum_{t=1}^{N} p_t \right) \ln \frac{1}{N} \sum_{t=1}^{N} p_t \tag{3.3}$$

Define weights g_t as $\partial S/\partial p_t$. Now to extremalize S with respect to variation in the filter, we set to zero

$$\frac{\partial S}{\partial f_j} = \sum_t \frac{\partial S}{\partial p_t} \frac{\partial p_t}{\partial f_j} = \sum_t g_t \frac{\partial p_t}{\partial f_j} = 2Y^T Gx \tag{3.4}$$

Notice the algebraic equivalence of setting (3.2) to zero with setting (3.4) to zero. The third example is equivalent to (3.2) and (3.4), but is more intuitive and computationally oriented. Consider a desire to scale up x_t when $g_t > \bar{g}$ and to scale down when $g_t < \bar{g}$. This thought can be expressed as

$$dx_t \approx (g_t - \bar{g}) x_t \tag{3.5}$$

In terms of the filter $x = Yf$ and $dx = Y\,df$ we have

$$Y\,df \approx (G - \bar{g}I)x \tag{3.6}$$

A traditional Levinson least-squares approach is to premultiply by Y^T:

$$Y^TY\,df = Y^T(G - \bar{g}I)x \tag{3.7}$$

These equations may be solved to find a df which may be used to update x. After many iterations we may achieve convergence. Convergence means that df comes out zero. And that, except for the constant $\bar{g}$, is equal to (3.2) and (3.4). Let us examine this constant. Premultiply (3.7) by f^T and assume convergence by assuming df = 0.

$$0 = (f^TY^T)(Y\,df) = (f^TY^T)(G - \bar{g}I)x \tag{3.8a}$$

$$0 = x^T\,dx = x^TGx - \bar{g}(x^Tx) \tag{3.8b}$$

$$\bar{g} = \frac{x^TGx}{x^Tx} \tag{3.8c}$$

From (3.8c) we see what the value of $\bar{g}$ must be in order for our assumption of convergence to be valid. It turns out that the "intrinsic entropies" always have $\bar{g} = 0$. Next, drop the convergence assumption but choose the value of $\bar{g}$ given by (3.8c). Inserting it into (3.8b), we see that dx is always perpendicular to x. This means that x preserves its length during iteration (to second order in Δx). So the $\bar{g}$ acts as a constant energy constraint for the extrinsic entropies, fixing up the G if need be. So the iterative procedure implied by (3.7) solves the zeroing of (3.2) and (3.4).

Although (3.7) determines the direction of df its magnitude, containing no second-derivative information about S, may be inappropriate. I believe a reasonable approach is to examine $dx = Y\,df$. If there are many sign agreements between the components of the dx of this iteration and those of the $\bar{dx}$ of the last iteration, then |dx| should be increased. In the event of

a preponderance of sign disagreements, |dx| should be decreased, say by about half.

Algorithm

$f = \delta(t)$

$x = y * f$

$\bar{dx} = x$

$\alpha = 1/10$ (α=1 might be faster if stable)

Begin iteration

p_t = slightly smoothed x_t^2 or its envelope

$g_t = \ln(p_t) - \ln \langle\langle p_t \rangle\rangle$ (for example)

$\bar{g} = x^T G x / x^T x$ (energy constraint)

$G = (g_t - \bar{g})I$

Levinson solve for df

$$(Y^T Y)df = Y^T G x$$

$dx = Y\, df$

$$\alpha = \alpha \frac{\text{signagree}\,(x,x) + 2\,\text{signagree}\,(dx,\bar{dx})}{2\,\text{signagree}\,(x,x)}$$

$x = x + dx\,\alpha$

$f = f + df\,\alpha$

$\bar{dx} = dx$

End iteration

Notice that as convergence occurs, $Y^T Gx$ will be tending toward zero, causing $\alpha\, dx$ to vanish regardless of whether or not α is vanishing. The real test of whether convergence is occurring is whether $||\alpha\, dx||/||x||$ is tending toward zero.

4. *Non-stationarity*: *Application to a Common-Shot Profile*

Reflection seismic data exhibits a severe decay of amplitude and change of color with time. Likewise there are systematic changes with offset. Is entropy processing ***dependent*** on a stationarity assumption? Is it ***sensitive*** to the familiar forms of non-stationarity? Must we devise ***ad hoc*** preprocessing to transform to stationarity, or can we directly allow for non-stationarity in our analysis for the entropy processing?

The gradient of the geometric inequality entropy is the inverse (square) envelope of the filter output plus a constant, $g = 1/\langle \bar{x}x \rangle + \text{const}$. The final equilibrium condition is $Y^T Gx = 0$ so the inverse envelope in the gradient is in the long range compensated for by the amplitude of y and the amplitude of x. In the short range, the envelope of x may be quite different from that of y, a fact which caused me to use numerous ***ad hoc*** stabilizing procedures during the descent. Also, the constant term, whose apparently modest purpose is to maintain normalization, is no longer gain-invariant when the $Y^T x$ is included. Realistically, we must conclude that in neither theory nor practice can entropy processing be regarded as ***gain-independent***. But what about sensitivity? Some quite good, as well as some quite poor, results have come out of experimental work where stationarity was largely ignored.

The gradient of extrinsic entropy defined by the parsimony inequality for a 2-D problem, say time $t = 1,2,\ldots,N_t$ and channel $j = 1,2,\ldots,N_j$, is

$$g_{jt} = \ln p_{jt} - \ln \frac{1}{N_j N_t} \sum_t \sum_j P_{jt} \tag{4.1}$$

If you suspected some variation in power from channel to channel you might prefer to define a different gradient for each channel, say

$$g_{jt} = \ln p_{jt} - \ln \frac{1}{N_t} \sum_t p_{jt} \tag{4.2}$$

Of course if $\sum_t p_{jt}$ really were independent of j then (4.2) would be algebraically equal to (4.1). In practice they will always differ, the advantage lying with (4.1) when they differ because of sample variance and the advantage lying with (4.2) when they differ because of systematic differences between channels. But would there be a practical difference between (4.1) and (4.2)? The practical difference clearly seen in the experimental work is that, if there is a systematic color difference between two channels, then the minimum entropy criterion will tend to choose filter color to amplify one channel and attenuate the other. This is clearly unacceptable. Wiggins told me that he preceded his entropy processing by conventional decon (which has the effect of balancing color). I was able to get away with ignoring this preprocess, but do not believe a successful production program can ignore it. So some kind of spectral balancing is necessary. But if we are admitting to systematic color and power variations from channel to channel we should surely admit to time variations within a channel. How does the ME filter choose its color? Perhaps the real role of the minimum entropy filter is to try to produce an output whose amplitude as a function of time is not correlated with its frequency as a function of time. Only the limited number of degrees of freedom prevent the final frequency from being completely uncorrelated with the amplitude.

Filter equations are invariant under simultaneous exponential scaling of data and filters. So we can always do a crude uniformization of power down the trace in this way, and we should always preprocess with such an exponential tilt. ***Gain distortion*** can be defined as geometric spreading and absorption correction, or automatic gain control. Gain distortion does seem to be a necessary part of transformation to stationarity. And transformation to stationarity seems to be relevant not only to estimating the debubble filter, but also to the estimation of the spectral balancing filters. At this point the reader may begin to fear that our data analysis is degenerating from scientific analysis to an ***ad hoc***, mushy, subjective, practitioner's art. Salvation comes from the fact that we need only apply gain distortion to filter ***outputs***. We need not apply it to filter ***inputs*** and therefore need not falsify our basic data or filter equations. We can obtain all the statistics we need from the gain-distorted filter outputs. Only at early stages in the iterative descent are statistics based on distorted raw data.

Let us assume that spectral balancing can be done before (hence in isolation from) debubble. Our first estimated balancing filters are just delta functions which means our first estimated balanced outputs are the raw data. Then we perform gain distortion on these outputs and find filter perturbations which will tend to balance the outputs. Apply the perturbed filters to the inputs and repeat. Rather than describe the spectral balancing in further detail (since iteration on it may be unimportant) let us go on to show the details of the debubble filter estimation procedure.

Recalling that x_t is the output of the debubble filter, we begin by stating our desire, namely something like

$$\frac{dx_t}{x_t} \approx \ln \left[\frac{\text{gain-corrected local envelope of } x_t}{\text{regional average of numerator}} \right] \tag{4.3}$$

To get more specific, we will need some definitions. Let

$1/t^2$	=	expected power in field trace, neglecting Q
$1/t$	=	expected power in slant stack, neglecting Q
x_t^2	=	square of trace
$\bar{x}x$	=	envelope of trace or smoothed local sum of squares
$\langle \bar{x}x \rangle$	=	locally smoothed envelope (say 10 ms = half-wavelength)
$\langle\langle \bar{x}x \rangle\rangle$	=	regionally smoother power (1 sec or so, longer than bubble)
$x^T x$	=	inner product, global sum of power

Now define a positive variable u_t which may be a fairly uniform function of time

$$u_t = \bar{x}x \frac{t}{\langle\langle \bar{x}xt \rangle\rangle} \tag{4.4}$$

As the denominator smoothing becomes very heavy then u becomes the envelope of the filtered seismogram corrected for spherical divergence. As the smoothing becomes very light then u becomes the envelope altered by "hard AGC." The sum of u over time may be expressed in matrix formalism as

$$\Sigma u = x^T D x = f^T Y^T D Y f \tag{4.5}$$

where D is diagonal matrix containing the scale factor of (4.4) which converts $\bar{x}x$ to u. Likewise,

$$\frac{1}{2}\frac{\partial}{\partial f}\Sigma u = Y^T D Y f = Y^T D x \qquad (4.6)$$

We want to push u away from uniformity by maximizing, say,

$$S = \frac{\Sigma u \ln u}{\Sigma u} - \ln \frac{1}{N}\Sigma u \qquad (4.7)$$

$$\frac{\partial S}{\partial u} = \ln u_t - \text{const}(t) \qquad (4.8)$$

Let G be a diagonal matrix with $\ln u_t$ on the diagonal. To maximize S with respect to variation of the filter f we set to zero

$$0 = \frac{\partial S}{\partial f} = \sum_t \frac{\partial S}{\partial u_t}\frac{\partial u_t}{\partial f} = Y^T D(G - \bar{g}I)x \qquad (4.9)$$

To achieve this result iteratively we start with (4.3) and substitute dx = Y df. Then using least squares, premultiply by $Y^T D$. Thus

$$(Y^T D Y)\, df = Y^T D(G - \bar{g}I)x \qquad (4.10)$$

As before, convergence will require $\bar{g} = x^T DGx/x^T Dx$, and we will find $x^T Dx$ constant during iteration because $x^T D\, dx = 0$. Iteratively solving this one may eventually find convergence, df = 0, when the right side, equaling (4.9), vanishes. The matrix $(Y^T DY)$ does not turn out to be Toeplitz. For computational convenience it could be approximated by a Toeplitz matrix made from gain-scaling the inputs y. Convergence implies the vanishing of the right-hand side no matter what (positive-definite) matrix lies on the left. So error of approximation on the left would be iterated away.

An economical, single-pass, approximate ME debubble filter can be achieved as follows. Take f to be a delta function, thus defining x, the gradient G, and the gain control D all from the process inputs y. Then solve (4.10) for df and quit with the filter f + df. In this case it might

be hazardous to approximate $Y^T DY$ as Toeplitz. But the effort in computing $Y^T DY$ involves both time sums and channel sums, so unless the filter f is extremely long, a conventional matrix inversion may not be costly compared to other costs. Also, weighted Levinson-type systems were considered in SEP-11, p. 167.

As a final item, let us incorporate the subjective opinion that information is not uniformly distributed down the time-axis of a seismogram. Both bandwidth and signal-to-noise ratio clearly diminish with time. I will suggest the subjective weighting factor $w = \exp(-t/3)$ as a conversion from the uniform variable u to the ***a priori*** information density. A sketch of the effect of this upon the main equations is

$$S = \frac{\sum w\, u \ln u}{\sum w\, u} - \ln \sum w\, u$$

$$\frac{\partial S}{\partial u} = w_t(\ln u_t + \text{const})$$

$$W^{1/2} Y\, df \approx W^{1/2} \ln(u)\, x$$

$$(Y^T WDY)\, df = Y^T WD(G - \bar{g}I)x$$

In conclusion, the theory of minimum entropy in the face of nonstationarity has advanced considerably since our last trials. It is now worthwhile to put together a production program to try to uncover any remaining misconceptions.

5. Seismogram Inversion

Seismogram inversion is the name generally given to the process of stripping a reflection seismogram of all its multiple reflections. The purpose is to enable comparison with well logs to allow extrapolation of geological conditions away from the well hole. Inversion of seismograms has ***not*** become a routine process such as filtering, deconvolution (spectral balancing), stacking, velocity analysis, or migration. It is not that the object is not

universally desired but that the process is as likely to destroy good data as to improve it. Anyway, the importance of the goal is evidenced by a continued popularity of effort, despite an erratic record of achievement. After my publications in this area in 1964 and 1968 I abandoned it, believing lateral variation (migration) to be more important in the analysis of field data. By 1973 there were a lot of deep water seismic data available that exhibited numerous obvious multiple reflections. To even the casual observer, it was apparent that the multiples were not being removed by efforts to predict them. Subtracting them out was not working nearly as well as stacking them out. This means that either multiples are incoherent, or we have bad predictive models. Undaunted, some doing inversion work have stressed the proper modeling of transmission coefficients and inner-bed multiples. Theoretically, such effects are second order small. Unfortunately, an objective test of the validity of such work is not as readily available as with seafloor multiples.

Pursuing my belief in the importance of lateral variation, Don C. Riley and I developed a theory for the incorporation of migration effects with multiples. Because of the need for vertical stacks the method turned out to be much more sensitive to the zero-offset data than to the further offsets. This disadvantage was largely overcome by later work on Snell stacks done with Raul Estevez. Much to my disappointment the process still did not lead to routine industrial application, despite additional development at Digicon, CGG, and elsewhere.

Present thoughts are that there is a remaining excess sensitivity to accurate knowledge of the seafloor reflection and the shot waveform. The difficulty is most severe in shallow water where high-order multiples are generated by repeated reflection from the seafloor. Thus a timing error of Δt on the seafloor becomes $n\Delta t$ on the n-th order multiple. Further, the seafloor must be convolved on itself n times to generate the multiple. If the original event lies between two samples on the time axis, it means that the interpolation operator must be convolved on itself that many times. This leads to considerable pulse broadening and dispersion. Even though a depth gauge gives a very accurate measure to the seafloor, the effective seismic depth and reflection strength will differ due to differing frequency content and the degree of consolidation of local sediments. Compounding all the above, the seafloor seismic reflection is often ***not recorded at those angles***

for which its multiples are a problem. Thus some process is needed to feed back information from high-order multiples to improve the quality of the seafloor reflection estimate. The whole issue is also associated with uncertainty in the source waveform. Given a successful method, it could add to the range of applicability of Riley and Estevez methods by enabling them to cope with that vast bulk of shallower water data where the cable "lead-in" cuts off the seafloor reflection for all angles of interest.

Using minimum entropy concepts we can hope to provide improved seafloor and shot-waveform estimates. The fact that minimum entropy has widely outdistanced least squares in shot-waveform estimation gives grounds for some optimism. Even more optimism is warranted by the fact that we seek not a whole, big, unknown waveform, but a small number of coefficients in the vicinity of the seafloor.

To see how this might be done, assume that a preliminary minimum entropy source wavelet has been deconvolved, but the t=0 position and the amplitude are still uncertain. For these we allocate a filter with about five adjustable coefficients. Likewise, we will use about a five-point operator to define the seafloor location. Take U and D to be up- and downgoing, spherical (or cylindrical) divergence corrected waves. For such a 1-D model, the reflection coefficient series C is given (to the "Noah" approximation) by the Z-transform ratio $C = U/D$. The free-surface reflection causes the downgoing wave D to be a superposition of the (residual) source S plus the negative of the upcoming wave U, so $D = S - U$. Let the inverse of S be the five-parameter filter F. Rearranging we see special mathematical justification for the term ***inversion***:

$$1 + C = 1 + \frac{U}{D} = \frac{D + U}{D} = \frac{S}{D} = \frac{1}{1 - U/S} = \frac{1}{1 - UF} \qquad (5.1)$$

By this definition the quantity 1+C is broad-band. It may be risky to apply a minimum entropy criterion to it unless it is first brought to a reasonable color. Suppose that a reasonable power spectrum is that of U. Let A be a minimum phase inverse (or symmetric) wavelet such that $\bar{A}A = 1/(\bar{U}U)$. Let us apply an entropy condition to X where

$$X = \frac{1 + C}{A} = \frac{1}{A(1 - UF)} \qquad (5.2)$$

The variation is

$$\delta X = \frac{1}{A(1 - UF)^2} (U\,\delta F + F\,\delta U)$$

$$= X^2 A(U\,\delta F + F\,\delta U) \qquad (5.3)$$

The data is the upcoming wave U and by δU we mean the variation that we intend to apply to the five coefficients in the digitized seismogram (slant stack or radial trace) nearest the seafloor. You can see that the color compensation filter A could be absorbed into δF and δU and had we not done so explicitly, it probably would have been there implicitly by the end of the iteration, so its precise choice should not be critical. The quantity X^2 is the convolution upon itself of our best reflection coefficient model. A "shifted column" matrix of X^2AU will be called Y_1. Likewise, a shifted column matrix of X^2AF will be called Y_2. With all this, (5.3) may be converted from Z-transform notation to matrix notation

$$\delta X = Y_1\,\delta F + Y_2\,\delta U \qquad (5.4)$$

In (5.4) δF and δU are five component vectors and δx has the length of a seismogram. Next we could haul out the machinery of multiple time-series analysis, but since we must proceed iteratively anyway, it is easier to let δF and δU be zero on alternate iterations. Setting, for example, δU equal to zero, and choosing a G from some entropy method, we get

$$\delta x = Gx \qquad (5.5)$$

$$Y_1\,\delta F \approx \delta x = Gx \qquad (5.6)$$

$$(Y_1^T Y_1)\,\delta F = Y_1^T Gx \qquad (5.7)$$

so we are now on the well-worn minimum entropy track. A difference is that after finding δF or δU we need to do the division (5.2). This calls for a little care if δF or δU gets too large. The sensitivity to error should

not be too great, however, because the coefficients of U are chosen zero before the seafloor, thereby limiting the growth of errors to the number of seafloor bounces.

Suppose the above process is successful. We are then in a position to incorporate transmission coefficients and inner bed multiples. This would not introduce any new degrees of freedom, it would only complicate the calculation of x and Y. Then we can find out whether inclusion of these physical factors leads to a consistent drop in entropy. This will be the first objective test of the Noah assumptions. Likewise it is an objective test of whether there is any utility in exploration of the theoretical phenomena of one-dimensional transmission coefficients and of inner bed multiples. That there can be any doubt about these arises from serious consideration of the nature of seismic reflection from real depositional sequences with real heterogeneity in all directions and knowledge that P-S conversions must occur in non-zero offset data.

6. ME Extrapolation and Power Spectrum

A classical problem in time-series analysis is the extrapolation of a finite segment of data off both its ends. Along with this question goes the much easier question (once the extrapolation has been done) of the determination of the energy spectrum of the extrapolated segment. In speaking of his famous maximum-entropy solution to this problem, John Burg once remarked that the reason for the high-resolution character of the method was widely misunderstood. The real reason for the great success of the method is that it is completely faithful to the observations, unlike windowing methods which effectively falsify the observations. That "entropy" (defined by the geometric inequality, see FGDP, p. 122) was maximized rather than maximizing almost anything else was a computational convenience, but, contrary to widespread opinion, was of no real significance.

To illustrate the validity of Burg's remarks I will show how we can get an excellent solution to the same problem by ***minimizing*** entropy instead of ***maximizing*** it. To ensure that the problem remains well posed, some apparently minor aspects of the problem formulation will be modified. In Burg's method, he fits a prediction filter of finite degree. The degree is necessarily

limited by the length of the data sample. In practice, the filter length is chosen by some poorly understood trade-off between resolution and sample variance. The implication of the finiteness of the degree is that the prediction off the end of the data sample goes asymptotically to zero. It does not become identically zero beyond some particular, finite distance. In the present problem formulation we will not use a prediction filter. Instead (on the basis of some vague resolution trade-off) we will choose a distance off the end of the data sample beyond which the extrapolation values all vanish. If we choose this distance to be too great, then we may expect to see sinusoidal continuation of the dominant frequency component of the data sample. If we choose the distance too small then we may expect to see the usual spectral broadening (smoothing) due to truncation. The appropriate distance for non-zero extrapolation values is deliberately left as a choice which lies outside the framework of the present theory. Which definition of entropy to minimize is also a matter for subjective choice. I, however, am biased in favor of a uniformly weighted, weak inequality. Maximize, say,

$$S'_{implicit} = \frac{\sum p_i \ln p_i}{\sum p_i} - \ln \frac{1}{N} \sum_{i=1}^{N} p_i \qquad (6.1)$$

where p_i is the power spectrum at the i-th frequency and N is as large as is computationally practical. The direction g_j in which to try to move against the constraints is

$$g_j = \ln p_j - \frac{\sum p \ln p}{\sum p} \cong \frac{\partial S}{\partial p_j} \qquad (6.2)$$

In words this says that where the log spectrum exceeds its average value we will try to boost the spectrum. Where it is less than the mean we will try to diminish it. Define

x = Column vector denoting the data sample.

$(x_\ell, x, x_r)^T$ = Data sample extended both to left and right.

$(0, x_\ell, x, x_r, 0)^T$ = Data sample on time domain of approximately infinite duration. Length of zero padding limited only by economics of calculation.

f = Fourier transform of above (column vector).

$[A_\ell\ B_\ell\ C\ B_r\ A_r]$ = Square matrix of Fourier transformation partitioned as data is partitioned.

G = Diagonal matrix with g_j on diagonal.

With these definitions the extended data sample x and its perturbation dx have the Fourier transformation $f + df$ where

$$[A_\ell\ B_\ell\ C\ B_r\ A_R] \left\{ \begin{bmatrix} 0 \\ x_\ell \\ x \\ x_r \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ dx_\ell \\ 0 \\ dx_r \\ 0 \end{bmatrix} \right\} = f + df \qquad (6.3)$$

Our basic desire is to change the complex Fourier transform f in magnitude so that

$$df \simeq G\, f \qquad (6.4)$$

From (6.3) this may be written

$$[B_\ell\ B_r] \begin{bmatrix} dx_\ell \\ dx_r \end{bmatrix} = df \simeq G\, f \qquad (6.5a)$$

or symbolically as

$$B\, dx \simeq G\, f \qquad (6.5b)$$

Think about the matrix B^T. It contains rows from the inverse Fourier transform matrix. Because of mutual orthogonality we have $B^T B = I$.

Premultiply (6.5b) by B^T to give a "least-squares" solution:

$$dx = B^T G f \tag{6.6}$$

These perturbations dx_ℓ and dx_r can be added into x_ℓ and x_r and the process can be iterated. Unfortunately, there is really nothing to be said about the uniqueness or certainty of convergence. The reputation of the method will have to be based on the subjective quality of the results for various examples. A favorable test case should be the sinc function. Can this method successfully extend the side lobes? I believe it will do so far more satisfactorily than Burg's method because of the fact that the sinc function has a box-car spectrum. Such a spectrum causes trouble for predictive methods but is ideal in this case. Another favorable test case would seem to be the sum of a few sinusoids.

Unfortunately there seems also to be no clever shortcut to the solution and the iteration might turn out to be a tedious one. At this stage I take the view that computers become cheaper every year while the cost of data collection rises. In more and more situations we are more concerned with the quality of results than with the underlying computer time used. Therefore, the computer algorithm which I will propose uses only the crudest estimate of the scale factor to be applied to the dx of equation (6.6). The basic idea is that if the dx of one iteration has many sign agreements with that of the previous iterations, say dx^-, then the step size is probably too small and it should be increased. Contrariwise, if dx has many sign disagreements with dx^- then the step size is probably too large and should be decreased. If the solutions turn out to be of exceptional subjective quality and economic utility we can always return to the task of finding a more clever descent procedure. The algorithm then is this:

```
Declare   real  α,x(N),dx(N),dx⁻(N),u(N),g(N)

          complex  f(N)
```

Initialize

x	=	$(0,0,x,0,0)$	Pad data with zeros.
α	=	1	Initialization.
$\bar{dx}$	=	arbitrary	Initialization.

Beginning of iteration loop

f	=	$FT(x)$	Fourier-transform extended data.
u	=	$f^{*}f$	Form spectrum.
g	=	$\ln u - (\Sigma\, u \ln u)/(\Sigma\, u)$	Form gradient weight.
dx	=	$FT^{-1}(gf)$	Inverse-transform weighted transform.
dx	=	$(0,dx_{\ell},0,dx_{r},0)$	Mask forbidden perturbations. (Now we have $B^{T}Gf$.)
NC	=	signagree $[(dx_{r},dx_{\ell}),(dx_{r},dx_{\ell})]$	Count free parameters.
α	=	$\alpha\left[\frac{1}{2} + \frac{3}{2\cdot NC}\ \text{signagree}(dx,\bar{dx})\right]$	Increase or decrease scale by as much as a factor of two.
x	=	$x + \alpha\, dx$	Update x_{l} and x_{r}.
$\bar{dx}$	=	dx	Save old direction.

Reiterate.

Convergence is occurring if some norm of $\alpha\, dx$ in diminishing in comparison with the norm of x. After convergence is achieved, the values being masked

off by the perturbations are of interest because they point to significant (or suspicious) data values.

The above problem, as posed, does not seem to have any very direct application to exploration geophysics. Nonetheless, if simple trials give encouraging results, there is a rather straightforward extension to what may be the ***most*** important problem in geophysical data analysis. That is, how should we extrapolate and interpolate spatially aliased data? It is known that human beings can do this with modest success, but I know of no satisfactory systematic approach to the problem. Clearly the question of spatially extending and interpolating a common-midpoint gather lies within this realm. To apply the present techniques, just consider the spectrum to be two-dimensional.

To shrink such a towering problem to manageable size, we might try the easier problem which results from a Fourier-transformed time axis. Now for each temporal frequency we have a (usually 48-point) function of space. The function may require extrapolation (as already described) and interpolation. The interpolation may be achieved by redefining the constraint mask to be a comb function (rather than zeros) over the geophone cable. Another application is the interpolation of data between parallel survey lines.

Satisfactory solutions to these practical problems will no doubt lead to better fundamental knowledge of properties of convexity inequalities and what makes some of them better than others. Also, better understanding of what constitutes a good answer may lead to better understanding of how we may converge more quickly, that is, how to quantify what it is we don't know and hence how to learn faster.

7. Convex Inequalities and Statistical Mechanics

A familiar inequality in science and engineering is the one which says that the mean of the squares exceeds the square of the mean, namely, the quadratic form Q is positive where

$$Q = \frac{1}{N} \sum_{i=1}^{N} x_i^2 - \left[\frac{1}{N} \sum_{i=1}^{N} x_i \right]^2 \geq 0 \qquad (7.1)$$

A slight generalization is a weighted sum where w_i are positive weights summing to unity, that is, given

$$w_i \geq 0 \quad \text{and} \quad \Sigma\, w_i = 1, \quad \text{then} \tag{7.2}$$

$$Q = \Sigma\, w_i x_i^2 - \left(\Sigma\, w_i x_i\right)^2 \geq 0 \tag{7.3}$$

This quadratic inequality is the simplest example of a whole family of inequalities. Any function $f(x)$ is said to be ***concave*** if $d^2f/dx^2 \geq 0$ and ***convex*** if $d^2f/dx^2 \leq 0$. For a ***concave*** function f and any weighted mean we can anticipate that

$$\text{mean}[f(x_i)] \geq f[\text{mean}(x_i)] \tag{7.4}$$

Equation (7.1) is the special case $f(x) = x^2$. Figure 7.1 illustrates the idea for a weighted mean of two numbers. The general proof of (4) is by induction successively including more numbers into the mean.

More interesting examples come from functions f whose second derivative changes sign at the origin. For such functions we still have convexity-type inequalities if we replace the arbitrary real numbers x_i by arbitrary positive numbers $r_i \geq 0$. A classic case is the ***harmonic*** inequality where $f(r) = 1/r$, $f''(r) = 1/r^3$, so

$$H = \Sigma\, \frac{w_i}{r_i} - \frac{1}{\Sigma\, w_i r_i} \geq 0 \tag{7.5}$$

A more important case is the ***geometric*** inequality. Here $f(r) = \ln(r)$, $f''(r) = -1/r^2$

$$G = \ln \Sigma\, w_i r_i - \Sigma\, w_i \ln r_i \geq 0 \tag{7.6a}$$

A somewhat more familiar form of the geometric inequality is found if we exponentiate both terms and choose weights $w_i = 1/N$.

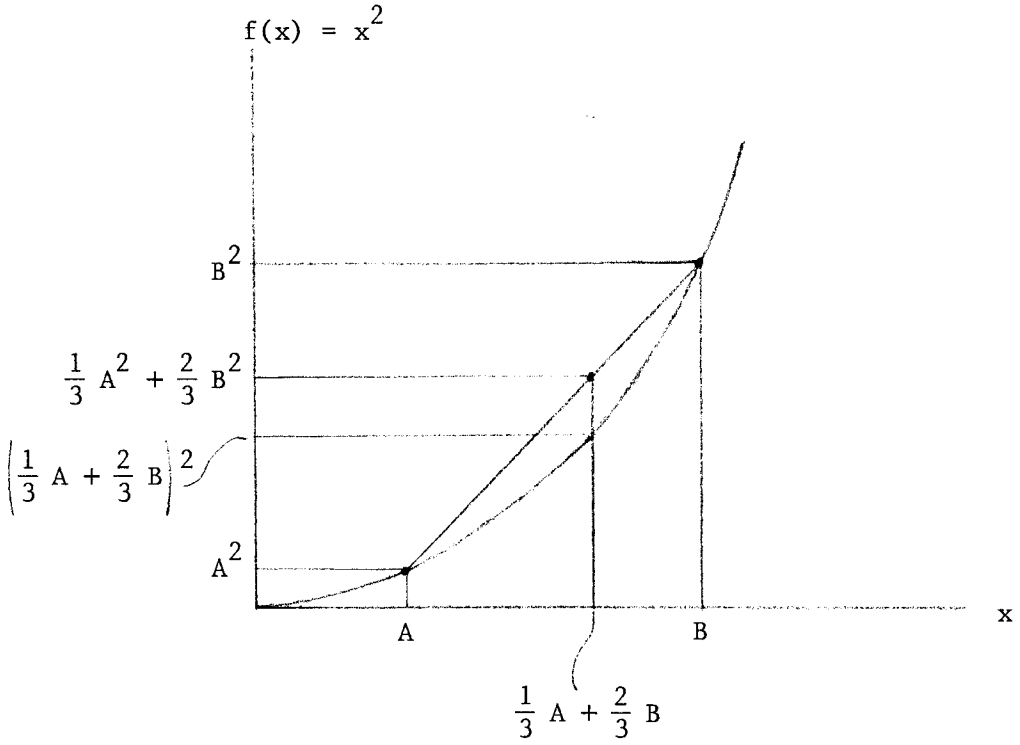

FIG. 7.1. Illustration that $\frac{1}{3}A^2 + \frac{2}{3}B^2 \geq \left(\frac{1}{3}A + \frac{2}{3}B\right)^2$

$$\frac{1}{N}\sum_{i=1}^{N} r_i \geq \prod_{i=1}^{N} r_i^{1/N} \tag{7.6b}$$

A most important inequality in information theory and thermodynamics is the one based upon $f(r) = r^{1+\epsilon}$ where ϵ tends to zero. I like to call this the ***weak*** inequality. A little bit of calculation enables us to go to the limit.

$$\sum w_i r_i^{1+\epsilon} \geq \left(\sum w_i r_i\right)^{1+\epsilon}$$

Take logarithms

$$\ln \sum w_i r_i^{1+\epsilon} \geq (1+\epsilon)\ln\left(\sum w_i r_i\right)$$

Expand both sides in a Taylor series. Note that the leading term cancels on both sides. Use

$$\frac{d}{dx} a^u = \ln(a)\, a^u \frac{du}{dx}$$

Divide both sides by ϵ and go to the limit $\epsilon=0$ getting

$$\frac{\Sigma\, w_i r_i \ln r_i}{\Sigma\, w_i r_i} \geq \ln \Sigma\, w_i r_i$$

We can now define a positive variable S' with or without a positive scaling factor Σ wr:

$$S'_{intrinsic} = \frac{\Sigma\, w_i r_i \ln r_i}{\Sigma\, w_i r_i} - \ln \Sigma\, w_i r_i \geq 0 \qquad (7.7a)$$

$$S'_{extrinsic} = \Sigma\, w_i r_i \ln r_i - (\Sigma\, w_i r_i) \ln(\Sigma\, w_i r_i) \geq 0 \qquad (7.7b)$$

These S' variables seem to relate to the negative of ***entropy*** in some (but not all) applications.

Recapitulating, we have defined positive statistics (Q,S',G,H) of a population r_i by using a convexity function f which is r to the power (2,1+ϵ,ϵ,-1), where ϵ tends to zero, namely (2,1,0,-1). Non-integral powers and other functions are also possible.

There are many curious properties of these statistics Q, S', G and H. Notice that they all vanish if the r_i are all equal to one another. Suppose an entire population r_i may be modified in some way by adjustment of some single scalar parameter X. Then we can say that choosing X to minimize Q, S', G or H is choosing X to drive the r_i toward ***uniformity***, ***homogeneity***, or ***equilibrium***. Likewise we may drive the r_i away from one another by adjusting X so that any of Q, S', G or H is maximized. Let us call the r_i ***microscopic*** variables. Let Q, S', G, H be called ***macroscopic-dependent*** variables and X a ***macroscopic-independent*** variable. Let us now more completely define the macroscopic-independent variable X by asserting that a change in X affects each and every r_i in the same way, namely such that

$$\frac{\partial r_i}{\partial X} = 1 \qquad (7.8)$$

Now let us see that the change of $S'_{extrinsic}$ with X gives the geometric inequality

$$S_{ext}' = \sum_i w_i r_i \ln r_i - \left(\sum_i w_i r_i\right) \ln\left(\sum_i w_i r_i\right) \tag{7.9}$$

$$\frac{\partial S'}{\partial r_j} = w_j \left[\ln r_j - \ln\left(\sum_i w_i r_i\right)\right] \tag{7.10}$$

$$\frac{\partial S'}{\partial X} = \sum_j \frac{\partial S'}{\partial r_j} \frac{\partial r_j}{\partial X} = -G \leq 0 \tag{7.11}$$

Next we see that the second derivative gives the harmonic inequality!

$$\frac{\partial^2 S'}{\partial X^2} = -\frac{\partial G}{\partial X} = \sum \frac{w_j}{r_j} - \frac{1}{\sum wr} = H \geq 0 \tag{7.12}$$

Let us now consider the physical application of the randomization of radiation trapped in a volume. At equilibrium we might imagine that the energy is more or less uniformly distributed throughout the volume. Imagine 10^{36} variables to describe this radiation and 10^{36}-1 constraining equations of motion. Now by selecting some dependent scalar macroscopic variable to minimize, such as Q, S', G, or H, we will be reducing the infinitude of solutions (actually only 10^{36} are mutually orthogonal) to a single solution. All of the macroscopic variables vanish if the energy distribution becomes completely homogeneous. However, the constraint equations may prevent complete homogeneity from being attained, say perhaps because of zero crossings of the wavelets, etc. Therefore it is a tricky business, subject to much theoretical and experimental analysis to decide what macroscopic variable to minimize. For definiteness let us try the assumption that we wish to minimize the ***weak*** inequality. Let us define

V	=	$\sum_{i=1}^{N} v_i$	Total volume is sum of volume of N cells.
U	=	$\sum_{i=1}^{N} u_i$	Total energy is sum of energy of each cell.
w_i	=	v_i/V	Weights are in proportion to cell size. Weights sum to unity even if volume V is varied.
r_i	=	u_i/v_i	Energy per unit volume tends to homogeneity at equilibrium.
$w_i r_i$	=	u_i/V	
$\sum w_i r_i$	=	U/V	
S	=	$-VS'_{extrinsic}$	Entropy S taken to be the negative of volume times the weak inequality.

Inserting into (7.7b) we have

$$S = U \ln\left(\frac{U}{V}\right) - \sum u_i \ln \frac{u_i}{v_i} \tag{7.13}$$

$$\frac{\partial S}{\partial u_i} = \ln \frac{U}{V} - \ln \frac{u_i}{v_i} \tag{7.14}$$

We define our macroscopic-independent variable X as U so that a change in the total energy U is shared by each cell in proportion to its volume. So

$$du_i = \frac{v_i}{V} dU \tag{7.15}$$

Multiplying (7.14) by (7.15) and summing, we get

$$\Delta S = \sum_i \frac{\partial S}{\partial u_i} du_i = G\Delta U \quad \text{or} \quad \frac{\partial S}{\partial U} = G \tag{7.16}$$

It is also easy to see that $(\partial^2 S)/(\partial U^2) = -H$. Likewise we define another macroscopic-independent variable as volume V so that a change in total volume V is shared by each cell in proportion to its volume. So

$$dv_i = \frac{v_i}{V} dV \tag{7.17}$$

$$\frac{\partial S}{\partial V} = \sum \frac{\partial S}{\partial v_i} \frac{\partial v_i}{\partial V} = 0 \tag{7.18}$$

Now let us copy some basic equations out of a standard thermodynamics book. These define temperature T, pressure P, and specific heat at constant volume c_V:

$$\left.\frac{\partial S}{\partial U}\right)_V = \frac{1}{T} \tag{7.19}$$

$$\left.\frac{\partial S}{\partial V}\right)_U = \frac{P}{T} \tag{7.20}$$

$$C_V = -\frac{\left(\frac{\partial S}{\partial U}\right)^2}{\frac{\partial^2 S}{\partial U^2}} \geq 0 \tag{7.21}$$

Inspection of (7.16) and (7.19) shows that high temperature is associated with $G \to 0$ and the u_i/v_i being very uniformly distributed. Likewise, low temperature is associated with $G \gg 0$ as would happen as one of the u_i tends to zero. Vanishing of (7.18), hence (7.20), means the entropy function we have chosen implies that pressure P vanishes, a reasonable enough idea since we are dealing with radiation. The positivity of specific heat in (7.21) is ensured by the harmonic inequality $H \geq 0$. The Nernst concept that specific heat vanishes at zero temperature is readily verified by letting one of the r_i tend to zero and seeing that $c_r = (\ln r_i)^2/(1/r_i)$ tends to zero.

Interesting as all this is, it does not prove that our original choice of entropy was correct, just that it does not seem to lead to any contradictions. An experimental procedure for determination of entropy is the integration of specific heat (see for example Callen, p. 326). A satisfactory mathematical basis for the field of thermodynamics seems to be available by combining the ideas found in the following two references:

(1) Callen, H. B., 1960, Thermodynamics: John Wiley & Sons, Inc.

(2) Weinhold, F., Metric geometry of equilibrium thermodynamics, J. Chemical Physics, v. 63, no. 6, p. 2479-2501.

But the subject at hand seems to be not so much ***thermodynamics*** as it is ***irreversible statistical mechanics***. I have been unable to discover a suitable reference for the mathematical superstructure of this field. But we seem to be off to a good start already and I would like to point to a few interesting operators and ideas.

A collection of interesting macroscopic-independent variables is given by

$$\frac{\partial}{\partial X} = \sum_j \frac{\partial}{\partial r_j} \tag{7.22a}$$

$$\frac{\partial}{\partial Y} = \sum_j r_j \frac{\partial}{\partial r_j} \tag{7.22b}$$

$$\frac{\partial}{\partial Z} = \sum_j \left[\frac{w_j}{r_j} - w_j \left(\sum_i \frac{w_i}{r_i}\right)\right] \frac{\partial}{\partial w_j} \tag{7.22c}$$

$$\frac{\partial}{\partial W} = \sum_j \left[w_j r_j - w_j \left(\sum w_i r_i\right)\right] \frac{\partial}{\partial w_i} \tag{7.22d}$$

Notice that the Z and W variables change the weights, but they do it in such a way that $\sum w_i$ is preserved equal to the constant 1.

One of the most interesting ideas is that of ***exact*** differential. It is disappointing to discover, for example, that for none of the above operators do we find the familiar thermodynamic expression

$$\frac{\partial}{\partial U}\frac{\partial S}{\partial V} = \frac{\partial}{\partial V}\frac{\partial S}{\partial U} \tag{7.23}$$

There are two reasons for this. The mathematical reason is that

$$\sum_k \frac{\partial r_k}{\partial X}\frac{\partial}{\partial r_k}\left[\sum_j \frac{\partial r_j}{\partial Y}\frac{\partial S}{\partial r_j}\right] \neq \sum_k \frac{\partial r_k}{\partial Y}\frac{\partial}{\partial r_k}\left[\sum_k \frac{\partial r_j}{\partial x}\frac{\partial S}{\partial r_j}\right] \tag{7.24}$$

And a physical reason is that entropy S can be an exact differential only in reversible thermodynamics, i.e. at equilibrium, and we have included no mathematical statement that we are describing equilibrium. The tantalizing prospect is that internal energy U should be an exact differential even in the non-equilibrium situation. Consider two volumes adjacent but isolated from one another. If they are suddenly interconnected by a physical or thermal leak, there would be a change of entropy but there would not be a change of internal energy. So, it seems that we should be able to find macroscopic-independent variables like those of (7.22a,b,c,d) such that

$$\left.\frac{\partial}{\partial S}\right]_V \left.\frac{\partial U}{\partial V}\right]_S = \left.\frac{\partial}{\partial V}\right]_S \left.\frac{\partial U}{\partial S}\right]_V \tag{7.25}$$

but I have been unable to find them.