(Don't let this notation mislead you into inverting the full matrix $\mathbf{W}(x) + \lambda\mathbf{S}$. You only need to solve for some $\mathbf{y}$ the linear system $(\mathbf{W}(x) + \lambda\mathbf{S}) \cdot \mathbf{y} = \mathbf{R}$, and then substitute $\mathbf{y}$ into both the numerators and denominators of 18.6.12 or 18.6.13.)

Equations (18.6.12) and (18.6.13) have a completely different character from the linearly regularized solutions to (18.5.7) and (18.5.8). The vectors and matrices in (18.6.12) all have size N, the number of measurements. There is no discretization of the underlying variable x, so M does not come into play at all. One solves a different $N \times N$ set of linear equations for each desired value of x. By contrast, in (18.5.8), one solves an $M \times M$ linear set, but only once. In general, the computational burden of repeatedly solving linear systems makes the Backus-Gilbert method unsuitable for other than one-dimensional problems.

How does one choose λ within the Backus-Gilbert scheme? As already mentioned, you can (in some cases *should*) make the choice *before* you see any actual data. For a given trial value of λ, and for a sequence of x's, use equation (18.6.12) to calculate $\mathbf{q}(x)$; then use equation (18.6.6) to plot the resolution functions $\widehat{\delta}(x, x')$ as a function of x'. These plots will exhibit the amplitude with which different underlying values x' contribute to the point $\widehat{u}(x)$ of your estimate. For the same value of λ, also plot the function $\sqrt{\text{Var}[\widehat{u}(x)]}$ using equation (18.6.8). (You need an estimate of your measurement covariance matrix for this.)

As you change λ you will see very explicitly the trade-off between resolution and stability. Pick the value that meets your needs. You can even choose λ to be a function of x, $\lambda = \lambda(x)$, in equations (18.6.12) and (18.6.13), should you desire to do so. (This is one benefit of solving a separate set of equations for each x.) For the chosen value or values of λ, you now have a quantitative understanding of your inverse solution procedure. This can prove invaluable if — once you are processing real data — you need to judge whether a particular feature, a spike or jump for example, is genuine, and/or is actually resolved. The Backus-Gilbert method has found particular success among geophysicists, who use it to obtain information about the structure of the Earth (e.g., density run with depth) from seismic travel time data.

CITED REFERENCES AND FURTHER READING:

Backus, G.E., and Gilbert, F. 1968, *Geophysical Journal of the Royal Astronomical Society*, vol. 16, pp. 169–205. [1]

Backus, G.E., and Gilbert, F. 1970, *Philosophical Transactions of the Royal Society of London A*, vol. 266, pp. 123–192. [2]

Parker, R.L. 1977, *Annual Review of Earth and Planetary Science*, vol. 5, pp. 35–64. [3]

Loredo, T.J., and Epstein, R.I. 1989, *Astrophysical Journal*, vol. 336, pp. 896–919. [4]

18.7 Maximum Entropy Image Restoration

Above, we commented that the association of certain inversion methods with Bayesian arguments is more historical accident than intellectual imperative. *Maximum entropy methods*, so-called, are notorious in this regard; to summarize these methods without some, at least introductory, Bayesian invocations would be to serve a steak without the sizzle, or a sundae without the cherry. We should also comment in passing that the connection between maximum entropy inversion methods, considered here, and maximum entropy spectral estimation, discussed in §13.7, is rather abstract. For practical purposes the two techniques, though both named *maximum entropy method* or *MEM*, are unrelated.

Bayes' Theorem, which follows from the standard axioms of probability, relates the conditional probabilities of two events, say A and B:

$$\text{Prob}(A|B) = \text{Prob}(A)\,\frac{\text{Prob}(B|A)}{\text{Prob}(B)} \tag{18.7.1}$$

Here Prob$(A|B)$ is the probability of A *given* that B has occurred, and similarly for Prob$(B|A)$, while Prob(A) and Prob(B) are unconditional probabilities.

"Bayesians" (so-called) adopt a broader interpretation of probabilities than do so-called "frequentists." To a Bayesian, $P(A|B)$ is a measure of the degree of plausibility of A (given B) on a scale ranging from zero to one. In this broader view, A and B need not be repeatable events; they can be propositions or hypotheses. The equations of probability theory then become a set of consistent rules for conducting inference [1,2]. Since plausibility is itself always conditioned on some, perhaps unarticulated, set of assumptions, all Bayesian probabilities are viewed as conditional on some collective background information I.

Suppose H is some hypothesis. Even before there exist any explicit data, a Bayesian can assign to H some degree of plausibility Prob$(H|I)$, called the "Bayesian prior." Now, when some data D_1 comes along, Bayes theorem tells how to reassess the plausibility of H,

$$\text{Prob}(H|D_1 I) = \text{Prob}(H|I)\frac{\text{Prob}(D_1|HI)}{\text{Prob}(D_1|I)} \tag{18.7.2}$$

The factor in the numerator on the right of equation (18.7.2) is calculable as the probability of a data set *given* the hypothesis (compare with "likelihood" in §15.1). The denominator, called the "prior predictive probability" of the data, is in this case merely a normalization constant which can be calculated by the requirement that the probability of all hypotheses should sum to unity. (In other Bayesian contexts, the prior predictive probabilities of two qualitatively different models can be used to assess their relative plausibility.)

If some additional data D_2 comes along tomorrow, we can further refine our estimate of H's probability, as

$$\text{Prob}(H|D_2 D_1 I) = \text{Prob}(H|D_1 I)\frac{\text{Prob}(D_2|HD_1 I)}{\text{Prob}(D_1|D_1 I)} \tag{18.7.3}$$

Using the product rule for probabilities, Prob$(AB|C) =$ Prob$(A|C)$Prob$(B|AC)$, we find that equations (18.7.2) and (18.7.3) imply

$$\text{Prob}(H|D_2 D_1 I) = \text{Prob}(H|I)\frac{\text{Prob}(D_2 D_1|HI)}{\text{Prob}(D_2 D_1|I)} \tag{18.7.4}$$

which shows that we would have gotten the same answer if all the data $D_1 D_2$ had been taken together.

From a Bayesian perspective, inverse problems are inference problems [3,4]. The underlying parameter set **u** is a hypothesis whose probability, given the measured data values **c**, and the Bayesian prior Prob($\mathbf{u}|I$) can be calculated. We might want to report a single "best" inverse **u**, the one that maximizes

$$\text{Prob}(\mathbf{u}|\mathbf{c}I) = \text{Prob}(\mathbf{c}|\mathbf{u}I)\frac{\text{Prob}(\mathbf{u}|I)}{\text{Prob}(\mathbf{c}|I)} \tag{18.7.5}$$

over all possible choices of **u**. Bayesian analysis also admits the possibility of reporting additional information that characterizes the region of possible **u**'s with high relative probability, the so-called "posterior bubble" in **u**.

The calculation of the probability of the data **c**, given the hypothesis **u** proceeds exactly as in the maximum likelihood method. For Gaussian errors, e.g., it is given by

$$\text{Prob}(\mathbf{c}|\mathbf{u}I) = \exp(-\frac{1}{2}\chi^2)\Delta u_1 \Delta u_2 \cdots \Delta u_M \tag{18.7.6}$$

where χ^2 is calculated from **u** and **c** using equation (18.4.9), and the Δu_μ's are constant, small ranges of the components of **u** whose actual magnitude is irrelevant, because they do not depend on **u** (compare equations 15.1.3 and 15.1.4).

In maximum likelihood estimation we, in effect, chose the prior Prob($\mathbf{u}|I$) to be constant. That was a luxury that we could afford when estimating a small number of parameters from a large amount of data. Here, the number of "parameters" (components of **u**) is comparable to or larger than the number of measured values (components of **c**); we *need* to have a nontrivial prior, Prob($\mathbf{u}|I$), to resolve the degeneracy of the solution.

In maximum entropy image restoration, that is where *entropy* comes in. The entropy of a physical system in some macroscopic state, usually denoted S, is the logarithm of the number of microscopically distinct configurations that all have the same macroscopic observables (i.e., consistent with the observed macroscopic state). Actually, we will find it useful to denote the *negative* of the entropy, also called the *negentropy*, by $H \equiv -S$ (a notation that goes back to Boltzmann). In situations where there is reason to believe that the *a priori* probabilities of the *microscopic* configurations are all the same (these situations are called *ergodic*), then the Bayesian prior Prob($\mathbf{u}|I$) for a *macroscopic* state with entropy S is proportional to $\exp(S)$ or $\exp(-H)$.

MEM uses this concept to assign a prior probability to any given underlying function **u**. For example [5-7], suppose that the measurement of luminance in each pixel is quantized to (in some units) an integer value. Let

$$U = \sum_{\mu=1}^{M} u_\mu \tag{18.7.7}$$

be the total number of luminance quanta in the whole image. Then we can base our "prior" on the notion that each luminance quantum has an equal *a priori* chance of being in any pixel. (See [8] for a more abstract justification of this idea.) The number of ways of getting a particular configuration **u** is

$$\frac{U!}{u_1!u_2!\cdots u_M!} \propto \exp\left[-\sum_\mu u_\mu \ln(u_\mu/U) + \frac{1}{2}\left(\ln U - \sum_\mu \ln u_\mu\right)\right] \tag{18.7.8}$$

Here the left side can be understood as the number of distinct orderings of all the luminance quanta, divided by the numbers of equivalent reorderings within each pixel, while the right side follows by Stirling's approximation to the factorial function. Taking the negative of the logarithm, and neglecting terms of order $\log U$ in the presence of terms of order U, we get the negentropy

$$H(\mathbf{u}) = \sum_{\mu=1}^{M} u_\mu \ln(u_\mu/U) \tag{18.7.9}$$

From equations (18.7.5), (18.7.6), and (18.7.9) we now seek to maximize

$$\text{Prob}(\mathbf{u}|\mathbf{c}) \propto \exp\left[-\frac{1}{2}\chi^2\right] \exp[-H(\mathbf{u})] \tag{18.7.10}$$

or, equivalently,

$$\text{minimize:} \quad -\ln\left[\,\text{Prob}(\mathbf{u}|\mathbf{c})\,\right] = \frac{1}{2}\chi^2[\mathbf{u}] + H(\mathbf{u}) = \frac{1}{2}\chi^2[\mathbf{u}] + \sum_{\mu=1}^{M} u_\mu \ln(u_\mu/U) \tag{18.7.11}$$

This ought to remind you of equation (18.4.11), or equation (18.5.6), or in fact any of our previous minimization principles along the lines of $\mathcal{A} + \lambda\mathcal{B}$, where $\lambda\mathcal{B} = H(\mathbf{u})$ is a regularizing operator. Where is λ? We need to put it in for exactly the reason discussed following equation (18.4.11): Degenerate inversions are likely to be able to achieve unrealistically small values of χ^2. We need an adjustable parameter to bring χ^2 into its expected narrow statistical range of $N \pm (2N)^{1/2}$. The discussion at the beginning of §18.4 showed that it makes no difference which term we attach the λ to. For consistency in notation, we absorb a factor 2 into λ and put it on the entropy term. (Another way to see the necessity of an undetermined λ factor is to note that it is necessary if our minimization principle is to be invariant under changing the units in which **u** is quantized, e.g., if an 8-bit analog-to-digital converter is replaced by a 12-bit one.) We can now also put "hats" back to indicate that this is the procedure for obtaining our chosen statistical estimator:

$$\text{minimize:} \quad \mathcal{A} + \lambda\mathcal{B} = \chi^2[\widehat{\mathbf{u}}] + \lambda H(\widehat{\mathbf{u}}) = \chi^2[\widehat{\mathbf{u}}] + \lambda \sum_{\mu=1}^{M} \widehat{u}_\mu \ln(\widehat{u}_\mu) \tag{18.7.12}$$

(Formally, we might also add a second Lagrange multiplier $\lambda' U$, to constrain the total intensity U to be constant.)

It is not hard to see that the negentropy, $H(\widehat{\mathbf{u}})$, is in fact a regularizing operator, similar to $\widehat{\mathbf{u}} \cdot \widehat{\mathbf{u}}$ (equation 18.4.11) or $\widehat{\mathbf{u}} \cdot \mathbf{H} \cdot \widehat{\mathbf{u}}$ (equation 18.5.6). The following of its properties are noteworthy:

1. When U is held constant, $H(\widehat{\mathbf{u}})$ is minimized for $\widehat{u}_\mu = U/M =$ constant, so it smooths in the sense of trying to achieve a constant solution, similar to equation (18.5.4). The fact that the constant solution is a minimum follows from the fact that the second derivative of $u \ln u$ is positive.

2. Unlike equation (18.5.4), however, $H(\widehat{\mathbf{u}})$ is *local*, in the sense that it does not difference neighboring pixels. It simply sums some function f, here

$$f(u) = u \ln u \tag{18.7.13}$$

over all pixels; it is invariant, in fact, under a complete scrambling of the pixels in an image. This form implies that $H(\widehat{\mathbf{u}})$ is not seriously increased by the occurrence of a small number of very bright pixels (point sources) embedded in a low-intensity smooth background.

3. $H(\widehat{\mathbf{u}})$ goes to infinite slope as any one pixel goes to zero. This causes it to enforce positivity of the image, without the necessity of additional deterministic constraints.
4. The biggest difference between $H(\widehat{\mathbf{u}})$ and the other regularizing operators that we have met is that $H(\widehat{\mathbf{u}})$ is not a quadratic functional of $\widehat{\mathbf{u}}$, so the equations obtained by varying equation (18.7.12) are *nonlinear*. This fact is itself worthy of some additional discussion.

Nonlinear equations are harder to solve than linear equations. For image processing, however, the large number of equations usually dictates an iterative solution procedure, even for linear equations, so the practical effect of the nonlinearity is somewhat mitigated. Below, we will summarize some of the methods that are successfully used for MEM inverse problems.

For some problems, notably the problem in radio-astronomy of image recovery from an incomplete set of Fourier coefficients, the superior performance of MEM inversion can be, in part, traced to the nonlinearity of $H(\widehat{\mathbf{u}})$. One way to see this [5] is to consider the limit of perfect measurements $\sigma_i \to 0$. In this case the χ^2 term in the minimization principle (18.7.12) gets replaced by a set of constraints, each with its own Lagrange multiplier, requiring agreement between model and data; that is,

$$\text{minimize:} \qquad \sum_j \lambda_j \left[c_j - \sum_\mu R_{j\mu} \widehat{u}_\mu \right] + H(\widehat{\mathbf{u}}) \tag{18.7.14}$$

(cf. equation 18.4.7). Setting the formal derivative with respect to $\widehat{u}_\mu$ to zero gives

$$\frac{\partial H}{\partial \widehat{u}_\mu} = f'(\widehat{u}_\mu) = \sum_j \lambda_j R_{j\mu} \tag{18.7.15}$$

or defining a function G as the inverse function of f',

$$\widehat{u}_\mu = G\left(\sum_j \lambda_j R_{j\mu} \right) \tag{18.7.16}$$

This solution is only formal, since the λ_j's must be found by requiring that equation (18.7.16) satisfy all the constraints built into equation (18.7.14). However, equation (18.7.16) does show the crucial fact that if G is *linear*, then the solution $\widehat{\mathbf{u}}$ contains *only* a linear combination of basis functions $R_{j\mu}$ corresponding to actual measurements j. This is equivalent to setting unmeasured c_j's to zero. Notice that the principal solution obtained from equation (18.4.11) in fact has a linear G.

In the problem of incomplete Fourier image reconstruction, the typical $R_{j\mu}$ has the form $\exp(-2\pi i \mathbf{k}_j \cdot \mathbf{x}_\mu)$, where $\mathbf{x}_\mu$ is a two-dimensional vector in the image space and $\mathbf{k}_\mu$ is a two-dimensional wave-vector. If an image contains strong point sources, then the effect of setting unmeasured c_j's to zero is to produce sidelobe ripples throughout the image plane. These ripples can mask any actual extended, low-intensity image features lying between the point sources. If, however, the slope of G is smaller for small values of its argument, larger for large values, then ripples in low-intensity portions of the image are relatively suppressed, while strong point sources will be relatively sharpened ("superresolution"). This behavior on the slope of G is equivalent to requiring $f'''(u) < 0$. For $f(u) = u \ln u$, we in fact have $f'''(u) = -1/u^2 < 0$.

In more picturesque language, the nonlinearity acts to "create" nonzero values for the unmeasured c_i's, so as to suppress the low-intensity ripple and sharpen the point sources.

Is MEM Really Magical?

How unique is the negentropy functional (18.7.9)? Recall that that equation is based on the assumption that luminance elements are *a priori* distributed over the pixels uniformly. If we instead had some other preferred *a priori* image in mind, one with pixel intensities m_μ, then it is easy to show that the negentropy becomes

$$H(\mathbf{u}) = \sum_{\mu=1}^{M} u_\mu \ln(u_\mu/m_\mu) + \text{constant} \tag{18.7.17}$$

(the constant can then be ignored). All the rest of the discussion then goes through.

More fundamentally, and despite statements by zealots to the contrary [7], there is actually nothing universal about the functional form $f(u) = u \ln u$. In some other physical situations (for example, the entropy of an electromagnetic field in the limit of many photons per mode, as in radio-astronomy) the physical negentropy functional is actually $f(u) = -\ln u$ (see [5] for other examples). In general, the question, "Entropy of what?" is not uniquely answerable in any particular situation. (See reference [9] for an attempt at articulating a more general principle that reduces to one or another entropy functional under appropriate circumstances.)

The four numbered properties summarized above, plus the desirable sign for nonlinearity, $f'''(u) < 0$, are all as true for $f(u) = -\ln u$ as for $f(u) = u \ln u$. In fact these properties are shared by a nonlinear function as simple as $f(u) = -\sqrt{u}$, which has no information theoretic justification at all (no logarithms!). MEM reconstructions of test images using any of these entropy forms are virtually indistinguishable [5].

By all available evidence, MEM seems to be neither more nor less than one usefully nonlinear version of the general regularization scheme $\mathcal{A} + \lambda\mathcal{B}$ that we have by now considered in many forms. Its peculiarities become strengths when applied to the reconstruction from incomplete Fourier data of images that are expected to be dominated by very bright point sources, but which also contain interesting low-intensity, extended sources. For images of some other character, there is no reason to suppose that MEM methods will generally dominate other regularization schemes, either ones already known or yet to be invented.

Algorithms for MEM

The goal is to find the vector $\widehat{\mathbf{u}}$ that minimizes $\mathcal{A} + \lambda\mathcal{B}$ where in the notation of equations (18.5.5), (18.5.6), and (18.7.13),

$$\mathcal{A} = |\mathbf{b} - \mathbf{A}\cdot\widehat{\mathbf{u}}|^2 \qquad \mathcal{B} = \sum_\mu f(\widehat{u}_\mu) \tag{18.7.18}$$

Compared with a "general" minimization problem, we have the advantage that we can compute the gradients and the second partial derivative matrices (Hessian matrices) explicitly,

$$\begin{aligned} \nabla\mathcal{A} &= 2(\mathbf{A}^T\cdot\mathbf{A}\cdot\widehat{\mathbf{u}} - \mathbf{A}^T\cdot\mathbf{b}) \qquad & \frac{\partial^2\mathcal{A}}{\partial\widehat{u}_\mu\partial\widehat{u}_\rho} &= [2\mathbf{A}^T\cdot\mathbf{A}]_{\mu\rho} \\ [\nabla\mathcal{B}]_\mu &= f'(\widehat{u}_\mu) & \frac{\partial^2\mathcal{B}}{\partial\widehat{u}_\mu\partial\widehat{u}_\rho} &= \delta_{\mu\rho}f''(\widehat{u}_\mu) \end{aligned} \tag{18.7.19}$$

It is important to note that while $\mathcal{A}$'s second partial derivative matrix cannot be stored (its size is the square of the number of pixels), it can be applied to any vector by first applying $\mathbf{A}$, then $\mathbf{A}^T$. In the case of reconstruction from incomplete Fourier data, or in the case of convolution with a translation invariant point spread function, these applications will typically involve several FFTs. Likewise, the calculation of the gradient $\nabla\mathcal{A}$ will involve FFTs in the application of $\mathbf{A}$ and $\mathbf{A}^T$.

While some success has been achieved with the classical conjugate gradient method (§10.6), it is often found that the nonlinearity in $f(u) = u\ln u$ causes problems. Attempted steps that give $\widehat{\mathbf{u}}$ with even one negative value must be cut in magnitude, sometimes so severely as to slow the solution to a crawl. The underlying problem is that the conjugate gradient method develops its information about the inverse of the Hessian matrix a bit at a time, while changing its location in the search space. When a nonlinear function is quite different from a pure quadratic form, the old information becomes obsolete before it gets usefully exploited.

Skilling and collaborators [6,7,10,11] developed a complicated but highly successful scheme, wherein a minimum is repeatedly sought not along a single search direction, but in a small- (typically three-) dimensional subspace, spanned by vectors that are calculated anew at each landing point. The subspace basis vectors are chosen in such a way as to avoid directions leading to negative values. One of the most successful choices is the three-dimensional subspace spanned by the vectors with components given by

$$\begin{aligned} e_\mu^{(1)} &= \widehat{u}_\mu[\nabla\mathcal{A}]_\mu \\ e_\mu^{(2)} &= \widehat{u}_\mu[\nabla\mathcal{B}]_\mu \\ e_\mu^{(3)} &= \frac{\widehat{u}_\mu\sum_\rho(\partial^2\mathcal{A}/\partial\widehat{u}_\mu\partial\widehat{u}_\rho)\widehat{u}_\rho[\nabla\mathcal{B}]_\rho}{\sqrt{\sum_\rho\widehat{u}_\rho\left([\nabla\mathcal{B}]_\rho\right)^2}} - \frac{\widehat{u}_\mu\sum_\rho(\partial^2\mathcal{A}/\partial\widehat{u}_\mu\partial\widehat{u}_\rho)\widehat{u}_\rho[\nabla\mathcal{A}]_\rho}{\sqrt{\sum_\rho\widehat{u}_\rho\left([\nabla\mathcal{A}]_\rho\right)^2}} \end{aligned} \tag{18.7.20}$$

(In these equations there is no sum over μ.) The form of the $\mathbf{e}^{(3)}$ has some justification if one views dot products as occurring in a space with the metric $g_{\mu\nu} = \delta_{\mu\nu}/u_\mu$, chosen to make zero values "far away"; see [6].

Within the three-dimensional subspace, the three-component gradient and nine-component Hessian matrix are computed by projection from the large space, and the minimum in the subspace is estimated by (trivially) solving three simultaneous linear equations, as in §10.7, equation (10.7.4). The size of a step $\Delta\widehat{\mathbf{u}}$ is required to be limited by the inequality

$$\sum_{\mu}(\Delta\widehat{u}_\mu)^2/\widehat{u}_\mu < (0.1 \text{ to } 0.5)U \tag{18.7.21}$$

Because the gradient directions $\nabla\mathcal{A}$ and $\nabla\mathcal{B}$ are separately available, it is possible to combine the minimum search with a simultaneous adjustment of λ so as finally to satisfy the desired constraint. There are various further tricks employed.

A less general, but in practice often equally satisfactory, approach is due to Cornwell and Evans [12]. Here, noting that $\mathcal{B}$'s Hessian (second partial derivative) matrix is diagonal, one asks whether there is a useful diagonal approximation to $\mathcal{A}$'s Hessian, namely $2\mathbf{A}^T\cdot\mathbf{A}$. If Λ_μ denotes the diagonal components of such an approximation, then a useful step in $\widehat{\mathbf{u}}$ would be

$$\Delta\widehat{u}_\mu = -\frac{1}{\Lambda_\mu + \lambda f''(\widehat{u}_\mu)}(\nabla\mathcal{A} + \lambda\nabla\mathcal{B}) \tag{18.7.22}$$

(again compare equation 10.7.4). Even more extreme, one might seek an approximation with constant diagonal elements, $\Lambda_\mu = \Lambda$, so that

$$\Delta\widehat{u}_\mu = -\frac{1}{\Lambda + \lambda f''(\widehat{u}_\mu)}(\nabla\mathcal{A} + \lambda\nabla\mathcal{B}) \tag{18.7.23}$$

Since $\mathbf{A}^T\cdot\mathbf{A}$ has something of the nature of a doubly convolved point spread function, and since in real cases one often has a point spread function with a sharp central peak, even the more extreme of these approximations is often fruitful. One starts with a rough estimate of Λ obtained from the $A_{i\mu}$'s, e.g.,

$$\Lambda \sim \left\langle \sum_i [A_{i\mu}]^2 \right\rangle \tag{18.7.24}$$

An accurate value is not important, since in practice Λ is adjusted adaptively: If Λ is too large, then equation (18.7.23)'s steps will be too small (that is, larger steps in the same direction will produce even greater decrease in $\mathcal{A}+\lambda\mathcal{B}$). If Λ is too small, then attempted steps will land in an unfeasible region (negative values of $\widehat{u}_\mu$), or will result in an increased $\mathcal{A}+\lambda\mathcal{B}$. There is an obvious similarity between the adjustment of Λ here and the Levenberg-Marquardt method of §15.5; this should not be too surprising, since MEM is closely akin to the problem of nonlinear least-squares fitting. Reference [12] also discusses how the value of $\Lambda + \lambda f''(\widehat{u}_\mu)$ can be used to adjust the Lagrange multiplier λ so as to converge to the desired value of χ^2.

All practical MEM algorithms are found to require on the order of 30 to 50 iterations to converge. This convergence behavior is not now understood in any fundamental way.

"Bayesian" versus "Historic" Maximum Entropy

Several more recent developments in maximum entropy image restoration go under the rubric "Bayesian" to distinguish them from the previous "historic" methods. See [13] for details and references.

- Better priors: We already noted that the entropy functional (equation 18.7.13) is invariant under scrambling all pixels and has no notion of smoothness. The so-called "intrinsic correlation function" (ICF) model (Ref. [13], where it is called "New MaxEnt") is similar enough to the entropy functional to allow similar algorithms, but it makes the values of neighboring pixels correlated, enforcing smoothness.
- Better estimation of λ: Above we chose λ to bring χ^2 into its expected narrow statistical range of $N \pm (2N)^{1/2}$. This in effect overestimates χ^2, however, since some effective number γ of parameters are being "fitted" in doing the reconstruction. A Bayesian approach leads to a self-consistent estimate of this γ and an objectively better choice for λ.

CITED REFERENCES AND FURTHER READING:

Jaynes, E.T. 1976, in *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, W.L. Harper and C.A. Hooker, eds. (Dordrecht: Reidel). [1]

Jaynes, E.T. 1985, in *Maximum-Entropy and Bayesian Methods in Inverse Problems*, C.R. Smith and W.T. Grandy, Jr., eds. (Dordrecht: Reidel). [2]

Jaynes, E.T. 1984, in *SIAM-AMS Proceedings*, vol. 14, D.W. McLaughlin, ed. (Providence, RI: American Mathematical Society). [3]

Titterington, D.M. 1985, *Astronomy and Astrophysics*, vol. 144, 381–387. [4]

Narayan, R., and Nityananda, R. 1986, *Annual Review of Astronomy and Astrophysics*, vol. 24, pp. 127–170. [5]

Skilling, J., and Bryan, R.K. 1984, *Monthly Notices of the Royal Astronomical Society*, vol. 211, pp. 111–124. [6]

Burch, S.F., Gull, S.F., and Skilling, J. 1983, *Computer Vision, Graphics and Image Processing*, vol. 23, pp. 113–128. [7]

Skilling, J. 1989, in *Maximum Entropy and Bayesian Methods*, J. Skilling, ed. (Boston: Kluwer). [8]

Frieden, B.R. 1983, *Journal of the Optical Society of America*, vol. 73, pp. 927–938. [9]

Skilling, J., and Gull, S.F. 1985, in *Maximum-Entropy and Bayesian Methods in Inverse Problems*, C.R. Smith and W.T. Grandy, Jr., eds. (Dordrecht: Reidel). [10]

Skilling, J. 1986, in *Maximum Entropy and Bayesian Methods in Applied Statistics*, J.H. Justice, ed. (Cambridge: Cambridge University Press). [11]

Cornwell, T.J., and Evans, K.F. 1985, *Astronomy and Astrophysics*, vol. 143, pp. 77–83. [12]

Gull, S.F. 1989, in *Maximum Entropy and Bayesian Methods*, J. Skilling, ed. (Boston: Kluwer). [13]