TA4 - 9:30

Proceedings of 24th Conference
on Decision and Control
Ft. Lauderdale, FL • December 1985

IMAGE RESTORATION AND SEGMENTATION USING THE ANNEALING ALGORITHM

E. D. Sontag[+] H. J. Sussmann[*]

Department of Mathematics
Rutgers University
New Brunswick, NJ 08903

We consider the problem of estimating a signal, which is known -- or assumed -- to be constant on each of the members $R_1,\ldots,R_m$ of a partition of a square lattice into m unknown regions, from the observation of the signal plus Gaussian noise. This is a nonlinear estimation problem, for which it is not appropriate to use the conditional expectation as the estimate. We show that, at least in principle, the "maximum likelihood estimator" (MLE) proposed by Geman and Geman lends itself to numerical computation using the annealing algorithm. We argue that the MLE by itself can be, under certain conditions (low signal to noise ratio), a very unsatisfactory estimator, in that it does worse than just deciding that the signal was zero. However, if combined with a rule which we propose, for deciding when to use and when to ignore it, the MLE can provide a reasonable suboptimal estimator. We then discuss preliminary numerical data obtained using the annealing method. These results indicate that: (a) the annealing algorithm performs remarkably well, and (b) a criterion can be formulated in terms of quantities *computed from the observed image* (without using a priori knowledge of the signal-to-noise ratio) for deciding when to keep the MLE.

§1. *The setting*.

For a positive integer n, we use Λ_n to denote the n by n integer lattice, i.e. the set of pairs (i,j) of integers between 1 and n. The members of Λ_n are called *sites*. The *nearest neighbors* of a site $s = (i,j)$ are the sites $s_1 = (i,j-1)$, $s_2 = (i-1,j)$, $s_3 = (i,j+1)$, $s_4 = (i+1,j)$. We write $s \sim s'$ to indicate that s and s' are nearest neighbors. The *diagonal neighbors* of s are the sites $s_1^d = (i-1,j-1)$, $s_2^d = (i-1,j+1)$, $s_3^d = (i+1,j+1)$, $s_4^d = (i+1,j-1)$. A *region* is a subset of Λ_n. An *edge* is a segment that joins two nearest neighbors. A region R is *connected* if any two points of R can be joined by a path which consists of edges joining pairs of points of R. A region R is *simply connected* if, whenever π is a simple loop in R (i.e. a simple closed path made of edges that join points of R), then all the points of Λ_n that are enclosed by π are in R. We will be particularly interested in regions that are both connected and simply connected. We will refer to them as *csc regions*. For a region R, we let $p(R)$ denote the number of points of R. We use $e(R)$ for the number of edges of R (i.e. of unordered pairs $(s,s') \in R\times R$ with $s \sim s'$). A *square* of R is a square whose four edges are edges of R. We use $s(R)$ to denote the number of squares of R.

Let R be a nonempty connected region. It can be proved that the R is simply connected if and only if

$$p(R)-e(R)+s(R) = 1. \qquad (1)$$

For any given n, let $\theta(n)$ denote the number of connected simply connected regions of the n by n lattice. We need the following bounds, communicated to us by M. Aizenman.

LEMMA 1. *There are constants* C_1, C_2 *such that* $1 < C_1 < C_2 < 2$ *and*

$$C_1^{n^2} \leq \theta(n) \leq C_2^{n^2} \quad \textit{for all} \quad n. \qquad (2)$$

Notice that 2^{n^2} is exactly the number of arbitrary subsets of the lattice. The upper bound of (2) is actually a bound on the number of connected regions. We omit the proof of (2), but it is important to remark that the lower bound depends on the existence of a large number of very "irregular" connected simply connected regions, whose boundaries are simple closed loops that nearly fill the whole lattice.

An m-*partition* is a finite sequence $P = (R_1,\ldots,R_m)$ of pairwise disjoint regions such that $R_1 \cup \ldots \cup R_m = \Lambda_n$. (Some of the R_i may be empty.) An m-partition is *connected* if all its regions are connected. A *simply connected* m-partition is defined similarly. (The empty set is connected and simply connected.) A *csc partition* is one where all the regions are connected and simply connected. A *csc partition with background* (cscb partition) is one where all the R_i are csc except possibly for R_1. We use the notations Part(m,n), $\mathrm{Part}_{csc}(m,n)$, $\mathrm{Part}_{cscb}(m,n)$ to denote, respectively, the class of all, all csc, all cscb m-partitions of Λ_n.

An *image* is a real-valued function on Λ_n. The set of all images, denoted by $\mathbb{R}^\Lambda$, is therefore an n^2-dimensional linear space. If P is an m-partition, then an image X is P-*constant* if it is constant on each region of P. For any m-partition P, let S_P denote the space of all P-constant images. Then S_P is a linear subspace of $\mathbb{R}^\Lambda$ of dimension $k(P)$, where $k(P)$ is the number of nonempty regions of P.

For any given m, let S^m denote the union of all the subspaces S_P, for all m-partitions P. Also, we let $S^{m,csc}$, $S^{m,cscb}$ denote the union of those S_P that correspond, respectively, to csc or cscb m-partitions. Clearly $S^{m,csc} \subseteq S^{m,cscb} \subseteq S^m$. As m

[+]Partially supported by US Air Force Grant AFOSR-85-0247.
[*]Partially supported by NSF Grant MCS78-02442-03.

varies, each of the three families of subsets of $\mathbb{R}^{\Lambda_n}$ defined above increases with m. Notice that the S^m, $S^{m,csc}$, $S^{m,cscb}$ are unions, and not spans, of linear subspaces, and so they are not themselves linear subspaces (except for $m = 1$ and $m = n^2$).

The estimation problem that we are interested in is as follows: given is an observed $Z \in \mathbb{R}^{\Lambda}$ which, for some small m, is thought to be of the form $X+U$, with U a Gaussian noise and X in one of the sets S^m, $S^{m,csc}$, $S^{m,cscb}$. The objective is to estimate X. To make this more precise, we must specify a probability distribution on the set of possible X's. The following is a natural choice: we suppose that (i) an integer K between 1 and m is chosen at random, so that the probability that $K = k$ is some given number p_k, (ii) having chosen K, we select a K-partition P at random, such that $k(P) = K$, (iii) having selected P, we let X be in S_P and distributed according to some density function on S_P.

The numbers p_k can be taken to be all equal, or - in the spirit of Statistical Mechanics - proportional to e^{-ck} for some $c > 0$. This latter choice penalizes partitions with too many regions. The particular partition P in step (ii) may be allowed to be arbitrary, or may be restricted to be csc or cscb; further, one may let all the allowed P have the same probability, or one can choose the probabilities so as to give higher weights to certain preferred regions. For instance, proceeding once again as in Statistical Mechanics, one can make the probability of P proportional to $e^{-\alpha|P|}$, where α is a constant and $|P|$ is the perimeter of P - defined simply as the number of pairs (s,s') of nearest neighbors for which s and s' are in different regions of P. (This has the effect of making very irregular regions less likely. However, this also tends to give too much weight to partitions where all but one of the regions are very small. One can counter this shortcoming by adding a priori constraints on the sizes of the regions, or by including extra exponential factors in the probabilities.)

Finally, once $P = (R_1,\ldots,R_K)$ has been selected choosing X amounts to choosing the constant values $c_1,\ldots,c_K$ that X is going to have on each of the regions. This can be done for instance by letting the c_i be uniformly distributed on some interval, or Gaussian.

As for the noise, we can take it to be Gaussian and white, or we can assume it to be correlated.

As an illustration, consider the case where $m = 2$ and the regions are unrestricted. Let X, U, Z be the image, noise and observation, respectively. Then X is specified by giving a partition $P = (R_1,R_2)$ and the values c_1, c_2 of X on R_1, R_2. The probability space is identified with the product $\text{Part}(2,n) \times \mathbb{R}^2 \times \mathbb{R}^{\Lambda_n}$, and the joint density of X and Z can be taken to be

$$p(P,c_1,c_2,y) = \frac{1}{N}e^{-H(P,c_1,c_2,y)} \tag{3}$$

where

$$H(P,c_1,c_2,y) = \frac{1}{2\sigma^2}(c_1^2+c_2^2) + \frac{1}{2\theta^2}\sum_{i=1}^{2}\sum_{s\in R_i}(y(s)-c_i)^2$$

$$+ \alpha \sum_{\substack{s\in R_1,s'\in R_2\\ s\sim s'}} 1 - \gamma\, \text{card}(R_1)\cdot\text{card}(R_2). \tag{4}$$

Here α, γ, σ, θ are constant, and N is a normalization constant. (This choice of p corresponds to taking $p_1 = 0$, $p_2 = 1$, i.e. not making a separate discrete decision whether to take $K = 1$ or 2. The constants σ, θ measure the power of the signal and the noise. The last two terms respectively penalize large perimeters - assuming $\alpha > 0$ - and partitions into regions of very unequal sizes - if $\gamma > 0$ - .) In practice, the computation of N is a nearly impossible task but, fortunately, the Metropolis algorithm (cf. [4]) makes it possible to generate samples of this distribution without having to know N.

The function H of (4) is called the Hamiltonian. The choice made in this example corresponds to taking c_1 and c_2 Gaussian, and letting the noise be uncorrelated. If, instead, we want to assume c_1 and c_2 to be uniformly distributed on some interval I, then we simply omit the first term, but restrict p and H to be defined only for c_1, c_2 in I. (Actually, H makes perfect sense even when I is the whole real line, i.e. when $\sigma^2 = +\infty$, but the probabilistic interpretation is lost because N becomes infinite.)

§2. The estimation problem.

It is well known that, no matter what X and Z are, the "best estimate" of X given Z is the conditional expectation $\mathbb{E}(X|Z)$. However, it is also well known that this estimator has some obvious drawbacks. For instance, because it is defined as an integral, the conditional expectation is often hard to compute. (In image processing problems, these integrals are sums over all possible configurations, and can seldom be evaluated.) An even more serious difficulty with conditional expectations is the following. Even if X is known to take values on some (nonconvex) subset S of a linear space, the values of $\mathbb{E}(X|Z)$ will in general fail themselves to be in S. (As a trivial illustration, consider the case of a random variable X, uniformly distributed on a sphere, U an independent Gaussian vector-valued random variable, and $Z = X+U$.) In our case, we encounter precisely this situation, because the sets S^m, $S^{m,csc}$, $S^{m,cscb}$ are clearly not convex, and we definitely want our estimate for X to lie in one of these sets.

A general methodology for dealing with estimation problems of this type in the image processing setting has been proposed by Geman and Geman [1]. Roughly, the method proposed involves four steps, namely: (1) model the images or objects to be detected as Gibbs states, i.e. let the probability of a configuration σ be proportional to $e^{-H(\sigma)}$, where H - the Hamiltonian - is a function on the set of all configurations; (2) make a model for the noise so that the joint probability of an image σ and observation σ_{obs} is given by $e^{-H(\sigma)-K(\sigma,\sigma^{obs})}$ for some function K, (3) let the "conditional Hamiltonian" be defined as $H_c(\sigma,\sigma^{obs}) = H(\sigma)+K(\sigma,\sigma^{obs})$, and use as the estimate for σ given σ^{obs} a value of σ which minimizes H_c (i.e. the "maximum likelihood estimator", henceforth abbreviated as MLE); finally (4) use the annealing algorithm to compute the above minimum.

Naturally, the justification of each of these steps raises a different set of questions. It is clear that, in our formulation of the model in §1, we have essentially followed the prescriptions of Steps 1 and 2, and we will not pursue the justification problem for these steps. Steps 3 and 4, however, require some further analysis, because some of the issues involved are particularly critical for our estimation problem. We begin by discussion Step 3. The problems regarding the use of the annealing algorithm will be touched upon later.

There is no general reason for believing that "maximum likelihood" estimators in the above sense are optimal or even reasonable. They clearly are so for linear Gaussian problems, where they happen to agree with the usual conditional expectations, but once nonlinearity or nongaussianness if allowed in they can have fairly undesirable properties. For instance, let X be a real random variable with density $N^{-1}e^{-\frac{\nu x^4}{4}+\frac{x^2}{2}}$, where $\nu > 0$ and N is a normalization constant. Let U be Gaussian $(0,1)$ and independent from X. Let $Z = X+U$. Let $\hat{X}$ be the "maximum likelihood" estimator for X given Y, i.e. $\hat{X} = \nu^{-1/3}Y^{1/3}$. Then a simple computation shows that, if ν is sufficiently large, then $\mathbb{E}(|\hat{X}-X|^2) > \mathbb{E}(|X|^2)$. This says that $\hat{X}$ is worse, in the mean square sense, than the trivial estimator X^* which consists of just taking $X^* = 0$ no matter what the observed value of Z is. For an even simpler example involving a discrete X, let X take values -1, 0, 1 with equal probability. Let U be Gaussian $(0,\sigma^2)$, and independent from X. Let $Z = X+U$. The MLE consists of estimating X to be the member of $\{-1,0,1\}$ which is closest to Z. If is large, the mean square error is worse than if we just estimate X to be zero.

Given that the MLE is not in general satisfactory, the question arises as to whether it is reasonable for our particular problem. At the moment we are not yet able to perform a complete rigorous analysis, but we will now discuss a model problem - which resembles our more complicated situations - where explicit calculations are possible.

Let n be a positive integer, and let X be an $\mathbb{R}^n$-valued random variable determined as follows. An integer $k \in \{1,\dots,n\}$ is chosen at random, with uniform probability, and then a value of the coordinate x_k is chosen with a normal $(0,\sigma^2)$ distribution, while the other coordinates are set equal to zero. So X is concentrated on $L = L_1 \cup \dots \cup L_n$, where L_i denotes the i-th axis, and its density is $(2\pi\sigma^2)^{-1/2}n^{-1}e^{-|x|^2/2\sigma^2}$. (Here $|x|^2 = \sum_{i=1}^{n} x_i^2$.) Let U be $\mathbb{R}^n$-valued Gaussian noise with variance θ^2, and independent from X. Let $Z = X+U$, and consider the problem of estimating X given Z. The joint density of X and Z, on the product $L \times \mathbb{R}^n$, is given by $e^{-h(x,z)}$ times a constant, where

$$h(x,z) = \frac{1}{2\sigma^2}|x|^2 + \frac{1}{2\theta^2}|z-x|^2. \tag{5}$$

The mean square optimal estimate $\hat{X}$ of X given Z is obtained by choosing, for each z, the vector $\hat{x} \in L$ that minimizes the integral

$$\int_L |\hat{x}-x|^2 e^{-h(x,z)} d_1x, \tag{6}$$

where we write d_1x to emphasize that (6) is a one-dimensional integral. An explicit computation shows that x is obtained as follows: first we find an index k such that $|z_k| = \max\{|z_j|: j = 1,\dots,n\}$. Then we set

$$\hat{x}_k = \rho\frac{w_k(z)z_k}{\sum_{j=1}^{n} w_j(z)}, \tag{7}$$

and we set all the other coordinates equal to zero. The weights w_j are given by

$$w_j(z) = e^{\alpha z_j^2}, \tag{8}$$

where $\alpha = \frac{\rho}{2\theta^2}$, $\rho = \frac{\sigma^2}{\sigma^2+\theta^2}$.

The MLE $\hat{X}_{ML}$ is even easier to compute. We get it by choosing k exactly as above, but then taking $(\hat{x}_{ML})_k = \rho z_k$. (Naturally, the other coordinates are again set equal to zero.)

Notice that both $\hat{X}$ and $\hat{X}_{ML}$ agree on the choice of k, and this choice does not depend on the values of σ and θ. That is, there is a certain amount of qualitative information (namely, which axis X was on) which is given as correctly by $\hat{X}_{ML}$ as by $\hat{X}$. Moreover, this information does not depend on the strength of the signal relative to the noise. On the other hand, $\hat{X}$ does something subtle that $\hat{X}_{ML}$ entirely misses: while $\hat{X}_{ML}$ takes $\hat{x}_k$ to be a definite proportion of z_k, $\hat{X}$ does look at the other z_j's in order to decide how much of z_k is going to be alotted to $\hat{x}_k$. If z_k is very large, but the other z_j's are very small, i.e. if z lies in a narrow cone about L_k, then $\hat{X}$ will assume that the observation comes primarily from the signal, and will estimate the signal to be almost as large as z. If, on the other hand, all the z_j are roughly equal, so that z is very far from any of the axes, then $\hat{X}$ will realize the noise must have been very large, and will therefore not take its own choice of k too seriously, and opt instead for guessing that the signal was very small, and therefore k was very uncertain anyhow. In the limiting case $\rho \to 1$, both $\hat{X}$ and $\hat{X}_{MLE}$ reduce to the obvious estimate $\hat{X} = Z$, which equals X, so that both $\hat{X}$ and $\hat{X}_{ML}$ have zero mean square error. In the other limiting case, as $\rho \to 0$, both $\hat{X}$ and $\hat{X}_{ML}$ go to 0. However, it can be proved that, when ρ is small enough, then $\mathbb{E}(|\hat{X}_{ML}-X|^2) > \mathbb{E}(|X|^2)$, i.e. $\hat{X}_{ML}$ is worse than just taking X to be zero. (This is of course not true for $\hat{X}$, because $\hat{X}$ is optimal, and so it is better than the zero estimator.) This suggests that there is a cutoff point $\bar{\rho}$ such that, for $\rho > \bar{\rho}$, $\hat{X}_{ML}$ is worth using in that, for instance, it does better than the zero estimator, but for $\rho < \bar{\rho}$ then $\hat{X}_{ML}$ should be ignored. If we want to estimate k, we can first use $\hat{X}_{ML}$ (i.e. find the axis to which z is closest) and then decide to accept this k or not based on whether $\rho > \overline{\rho}$ or $\rho < \overline{\rho}$. The first of the two decisions clearly does not depend on knowing ρ. It would be desirable if the second one could also be made independently of ρ. For this, a Bayesian approach would be needed to estimate ρ from the observed Z. On the other hand, it is intuitively clear that, if a value z has been observed for the random variable Z which happens to be very close to one of the axes, then this is very unlikely to have happened because of the noise, and one would tend to believe that ρ was large. Similarly, if many z_j's are large, then ρ is more likely to have been small.

We now argue heuristically and attempt to show that something like the preceding analysis should apply to our original estimation problem. We consider the case $m = 2$ and take a Hamiltonian of the form (4), with $\alpha = \gamma = 0$. Moreover, we assume that all images are first normalized by subtracting their mean over the lattice, so that all images have mean zero and there are no nontrivial constant images. Then each partition $\mathcal{P} \in \text{Part}(2,n)$ gives rise to a <u>one-dimensional</u> subspace $S_{\mathcal{P}}$ of the space of all images. The analogy with the above simplified model is apparent, except for the fact that we are now dealing with a more general collection Σ_n of subspaces than the set of axes. Moreover, Σ_n is actually much larger, since the ambient space has dimension n^2-1 but the number of partitions behaves like e^{cn^2} for some $c > 0$, as $n \to \infty$. Let us assume that the analysis of our model problem still applies here. The MLE is obviously (and rigorously) obtained by first finding the subspace $S_{\mathcal{P}}$ to which the observed image Z is closest and then finding the constant values of X on the regions of $\mathcal{P}$. The partition $\mathcal{P} = (\mathcal{R}_1,\mathcal{R}_2)$ is, simply, the one such that, if we take $\tilde{c}_1$, $\tilde{c}_2$ to be the means of Z on $\mathcal{R}_1$ and $\mathcal{R}_2$ respectively, then the image with values $\tilde{c}_i$ on the $\mathcal{R}_i$ will best approximate Z in the mean square sense. Then the constants c_1, c_2 are taken to be multiples of $\tilde{c}_1$, $\tilde{c}_2$, by the factor ρ, where $\rho = \sigma^2/\sigma^2+\theta^2$. The optimum estimator, if we believe the analogy with the model problem, would be given exactly like the MLE, except that the $\tilde{c}_i$ would have to be multiplied by an extra factor. However, this factor would involve, in its denominator, a sum of terms $e^{c\pi_Q(Z)^2}$ where, for each partition Q, $\pi_Q(Z)$ denotes the norm of the orthogonal projection of Z on S_Q. Since the sum would be over all partitions, it is clear that computing the optimum estimator is likely to be much harder than computing the MLE. On the other hand, we know that the MLE is an overoptimistic estimator, in that it gives a multiple of the optimum estimator by a factor > 1. Hence it should be possible to improve upon the MLE by making a rough estimate of the extra factor. To this effect, we propose the following heuristic argument. Suppose we have k one-dimensional subspaces of $\mathbb{R}^p$, where p is very large, and k grows with p. Suppose the subspaces are "uniformly distributed". One can then estimate the area of the piece of the $(p-1)$-dimensional unit sphere that consists of those points that are closest to one particular subspace. This area is roughly equal to A_p/k, where A_p is the area of the $(p-1)$-dimensional unit sphere. Since $A_p = \dfrac{2\pi^{p/2}}{\Gamma(\frac{p}{2})}$, we see that the area equals $\dfrac{2\pi^{p/2}}{k\Gamma(\frac{p}{2})}$. If this piece of sphere is actually a $(p-1)$-dimensional disc of radius r, then its volume is $\dfrac{r^{p-1}}{p} A_{p-1}$, i.e. $\dfrac{2r^{p-1}\pi^{\frac{p-1}{2}}}{p\Gamma(\frac{p-1}{2})}$. So r roughly equals $1/k^{1/p}$. As $p \to \infty$, r will stay bounded away from 0 and 1 if k goes like C^p for some $C > 1$, and then r will approach $1/C$. In our estimation problem, p is n^2, and k is the number of partitions. For the case of unrestricted partitions, $k = 2^{n^2}$, and so $r \sim \frac{1}{2}$. This says that about $\frac{3}{4}$ of the sample variance of a typical purely random image should appear to be explainable by assuming that the image is $\mathcal{P}$-constant for some partition $\mathcal{P}$. For csc partitions a similar conclusion should hold, because Lemma 1 tells us that k also behaves like an exponential of n^2. The constant C is smaller, however, so that there will still be a definite, but smaller, fraction of the variance that will appear to be explainable by approximating by a signal that is constant on a 2-partition. (For unrestricted partitions, one can obtain directly a better estimate, and show that the fraction of the variance that will be accounted for by a 2-constant approximation is $\frac{2}{\pi}$. This just follows from the fact that $\mathbb{E}(|X|) = \sqrt{2/\pi}$ if X is normal $(0,1)$.) So, if we compute the best approximating 2-partition $\mathcal{P}$, we will typically reduce the variance by a fixed factor, and when this happens no particular significance should be attached to $\mathcal{P}$. However, if a much larger reduction results, then it is likely to mean that the power of the signal relative to the noise was large, and $\mathcal{P}$ really tells us something about the signal.

The preceding heuristic argument also suggests that, if we were to choose smaller classes of partitions, whose sizes grow less than exponentially in n^2, then for large n no significant reduction of the variance would be achieved for a typical random image, and any observed reduction would be significant.

Thus, we suggest the following procedure. First compute the MLE, but then estimate the signal as follows: let Y be the image obtained from the MLE, so that $Y = \rho Y^*$, where Y^* is the orthogonal projection of the observation Z on some space $S_{\mathcal{P}}$. If ρ is larger than a certain $\overline{\rho}$, then keep Y. Otherwise, just estimate X to be zero. If σ, θ are not known, just take $\sigma = \infty$ in (4), and compute the MLE, which will be equal to Y^*. Then determine the ratio $\dfrac{\|Z-Y^*\|^2}{\|Z\|^2}$. If this ratio is significantly smaller than some number x (which should be about equal to 0.34 for the unrestricted case, and larger for the csc case), then regard $\mathcal{P}$ as a good estimate of the partition for the original signal. If the ratio is larger than x, estimate X to be zero.

§3. <u>Computing the MLE by the annealing method</u>.

The numerical calculation of the MLE is possible, at least in principle, using the annealing algorithm. To see this we must show: (a) that the classes of partitions considered here are "connected by switchings", in the sense that we can connect any two partitions in the class by means of a sequence of switchings at single sites, without ever leaving the class, (b) that the possible switchings that can be made at a site can be recognized by purely local considerations, and (c) that the change in the value of the Hamiltonian can be calculated locally. (Strictly speaking, (b) and (c) are not necessary, since the annealing procedure is in principle applicable to minimization problems on arbitrary finite sets. However, in practice, for image problems, one needs (b) and (c) if one is to avoid having to carry out searches over the whole lattice at each basic iteration step.) The truth of (a) and (b) is completely trivial for unrestricted partitions.

For csc partitions more work is required, but the conclusion is the same. (For example, to decide whether or not one can switch a site s from Region 1 to Region 2, one has to make sure that adding s to 2 will keep 2 a csc region. Checking whether 2 remains connected is easy, for this will happen if and only if 2 was empty or at least one nearest neighbor of s was already in it. Once this is done, the preservation of simply connectedness is equivalent to the condition that the left side of (1) does not change, i.e. that adding s to Region 2 adds to it exactly one more edge than it adds squares.)

The verification of (c) is quite easy, provided that we allow ourselves to carry in the iterations the size of each region and the sum of Z over each region. (These numbers are easily updated at each basic iteration by purely local computations.)

Once we know that annealing is in principle a possible method for computing P, we must consider whether the computation actually works in practice, i.e. whether the method will take us reasonably close to the minimum in a realistic number of iterations. The existing theorems on the convergence of the annealing algorithm (cf., e.g., [2], [3]) do not suffice to establish this since, when the estimates obtained in these theorems are applied to our problem, they give a number of iterations that behaves like e^{cn^2}, $c > 0$. This only shows that annealing is not worse than exhaustive search, which is certainly a necessary but by no means sufficient condition for an optimization method to be worth using. In the absence of sharper theoretical bounds, we can only attempt to answer the question empirically. As we shall see, the results we have obtained strongly suggest that, for our estimation problem, the annealing algorithm works very well.

§4. Numerical results.

We run 25 trials, each consisting of: (a) generation of a partition P into two regions (with Region 2 being csc) using the Metropolis algorithm, (b) generation of a P-constant image X, by choosing each of the two values of X to be Gaussian $(0,\sigma^2)$, (c) generation of an "observed image" Z, by adding uncorrelated Gaussian noise to X, (d) two applications of the annealing algorithms to compute the MLE, one with "restricted switchings" (i.e. keeping Region 2 csc) and the other one unrestricted. The lattice was 28×28. The ratio $\sigma^2/\sigma^2+\theta^2$ was made to vary from 0 to 1. The Hamiltonian used in the annealing computation was that of (4), with $\alpha = \gamma = 0$. The first term of the right side of (4) was omitted, since this term has no effect on the determination of the partition. Because of this, together with the fact that $\alpha = \gamma = 0$, we obtain a situation where, for the unrestricted case, the optimization problem is in fact reducible to a clustering problem, and can also be solved algorithmically in time $O(n^2 \log n)$. Moreover, when $\rho = \sigma^2/\sigma^2+\theta^2$ is small, we can estimate a priori that $\|Z-Y^*\|$ should typically be about $(1-\frac{2}{\pi})\|Z\|$. Thus we get an independent test to see whether the annealing algorithm is really computing the minimum.

The results show that, for small ρ, and unrestricted switchings, the optimum computed by annealing was very close to the predicted value. This suggests that, for the problems considered here, annealing is actually computing the optimum. Convergence was quite fast: after 120 iterations (passes through the lattice) no significant reduction in the value of h was detected. Let $\lambda_r = \frac{\|Z-Y^*\|^2}{\|Z\|^2}$, $\mu_r = \frac{\|Y^*-X\|^2}{\|X\|^2}$, for Y^* computed using restricted switchings. Let λ_u, μ_u be defined similarly, using unrestricted switchings. A plot of λ_u against ρ shows that, for low values of ρ, λ_u is constant and its value is about 0.3. At $\rho = \rho_u^{\#} \sim 0.6$, $\lambda_u(\rho)$ becomes significantly smaller, and decreases to zero as $\rho \to 1$. The behavior of λ_r appears to be similar, except that (a) the constant value for small ρ is about 0.55, i.e. significantly larger than the value for the unrestricted case, in good agreement with our theoretical discussion, and (b) the cutoff value $\rho_r^{\#}$ is probably lower, $\rho_r^{\#} \sim 0.5$. (Since λ_r fluctuates more than λ_u, the determination of $\rho_r^{\#}$ is more uncertain.) Both μ_r and μ_u decrease steadily as functions of ρ. The graphs go through the value 1 at about $\rho \sim 0.4$. This means that, for $\rho < 0.4$, the MLE is worse than just taking $\hat{X} = 0$. For $\rho > 0.4$, the MLE does better. If we do not know ρ, we can get partial information by computing λ_u (or λ_r). If λ_u is significantly less than 0.3 (or $\lambda_r < 0.55$), then this says that ρ is large enough that the MLE must be taken seriously, and used as an estimator for the signal.

We also run a series of trials in which the signal was simply a dark square on a light background. In this case, a similar behavior was observed. However, another interesting phenomenon was noticed for the case of very low noise. In some trials with restricted switchings (i.e. with Region 2 - but not Region 1 - required to be csc), the algorithm several times became fixated on taking Region 2 to be the outer one and Region 1 to be the square. Since Region 2 was required to be csc, the end result was a square of ones on a background of twos, except that there was a string of ones joining the square to the boundary. Such a configuration is clearly a local minimum which is not global. However, this minimum is very close in value to the global one, although very far in configuration space. Going from this local minimum to the global one would have required climbing a very high hill, and the algorithm was unable to do it.

§5. Conclusion.

Our preliminary data indicate that, for the estimation problem considered here, the annealing algorithm works rather well to get quite close to the minimum in a reasonably small number of iterations. However, the phenomenon of "getting trapped at a local minimum" can occur, especially for restricted switchings. (This shows that the imposition of constraints on configurations, in order to obtain configurations with special properties, also has the undesirable side effect of making motion in configuration space more difficult.)

As for the question of how the computed MLE can be used to estimated the signal, the answer is more complicated, as was shown before, but there clearly is a range of values of ρ such that (a) one can assess from the data whether ρ is in that range, and (b) if ρ is in that range, then the MLE is good.

References

1. Geman, S. and D. Geman: Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images, IEEE Transactions, PAMI 6 (1984), 721-741.

2. Gidas, B.: Nonstationary Markov Chains and Convergence of the Annealing Algorithm, J. Statistical Physics 39 (1985), 73-131.

3. Hajek, B.: Cooling Schedules for Optimal Annealing, preprint, 1985.

4. Metropolis, N., et al.: Equations of State Calculations by Fast Computing Machines, J. Chem. Phys. 21 (1953), 1087-1091.