A Comparison Between MFA and GNC

Griff L. Bilbro

Center for Communications and Signal Processing
Department Electrical and Computer Engineering
North Carolina State University

TR-90/11
September 1990

A Comparison between MFA and GNC

Griff L. Bilbro

September 26, 1990

1 Introduction

We will show that Mean Field Annealing (MFA) provides a unified approach to several difficult optimization problems. All the resulting MFA algorithms resemble the Graduated Nonconvexity Algorithm (GNC) developed by Blake and Zisserman[4] to solve certain image processing problems. MFA is based on Simulated Annealing (SA) and derives its power and generality from that popular optimization procedure. MFA differs from SA by analytically approximating the relevant Gibbs distribution rather than stochastically simulating it.

Simulated Annealing is slow but is otherwise simple, general, easy to apply to new problems, and remarkably successful even when theoretical conditions for convergence are not met[10]. SA works by gradually cooling an on-going stochastic simulation of a Gibbs distribution. Mean field theory provides a deterministic approximation to a Gibbs distribution which also can be cooled in the same way to produce a Mean Field Annealing (MFA) algorithm. Many SA algorithms can be converted to analogous MFA algorithms that run in 1/50 the time required by the SA version[9, 1, 8, 2]. However because it is an approximation, MFA does not inherit any guarantee of convergence even when the analogous SA does converge.

The mean field is usually restricted to Ising-like systems described by an energy function (or Hamiltonian) involving binary variables $s = \{s_i\}_{i=1}^{i=N}$

$$H[s] = -\sum_i h_i s_i - \sum_{ij} v_{ij} s_i s_j, \tag{1}$$

but recently we have extended MFA to a wider class of problems. This extension depends on Peierls' inequality $F \leq W$ which bounds the exact free energy at temperature T of a system of s described by Equation 1

$$F = -T \ln \sum_{\{s\}} \exp(-H/T), \tag{2}$$

by the Weiss free energy W

$$W = F_0 + \langle H - H_0 \rangle. \tag{3}$$

Here H_0 is an arbitrary function of s. The average is with respect to the density $Z_0^{-1} \exp(-H_0/T)$ normalized over all possible configurations of the s by the factor

$$Z_0 = \sum_s \exp(-H_0/T) \tag{4}$$

which is also used to define $F_0 = -T \ln Z_0$.

The free energy F function characterizes the equilibrium of the system described by Equation 1 at temperature T. The utility of the bound $F \leq W$ is that it remains valid even if H_0 depends on adjustable parameters in addition to s. It can be shown[3] that equality obtains if and only if $H_0 = H$. In this mean field sense, adjusting H_0 to minimize W adjusts H_0 to most closely resemble H. Even Peierls' inequality was previously restricted to Ising-like problems, but we have recently shown how to choose H_0 to treat continuous variables as well[8, 2]. Peierls' inequality was originally derived in the context of statistical mechanics, but we have recently constructed an information-theoretic derivation[3].

MFA converts an optimization problem into the limiting member of a *family* of optimization problems. Instead of directly varying H, MFA varies a certain weighted average of H. The width of averaging kernel depends on a scalar called the temperature T. For small T the kernel approaches a Dirac delta function and in the limit the original problem is recovered. For large T fine structure of the original problem is averaged away and in this limit the objective becomes convex even if the original objective was not. MFA works when the large-T optimum approaches the best (or at least a good) low-T optimum as T is reduced.

Geiger and Girosi first reported a relation between MFA and GNC optimization of the weak membrane. They showed that MFA can be applied to the weak membrane to obtain an algorithm that is qualitatively identical to GNC for the same problem[6]. In this report we will show that MFA generally produces "GNC-like" algorithms for several optimization problems.

2 A binary problem

The H of Equation 1 is useful for graph bisection if v_{ij} is the adjacency matrix, h_i reflects some externally specified preference for assigning node i to a particular partition, and the value of s_i determines the assignment of node i[1, 5]. A slightly more complicated form of H has been used to restore binary images[3]. Here consider the simplest one-dimensional case where $h_i = h, \forall i$ and all v vanish except $v_{ij} = v, \forall j = i+1$ so that

$$H = -h \sum_{i=1}^{i=N} s_i - v \sum_{i=1}^{i=N} s_i s_{i+1} \quad (5)$$

with the s connected in a ring $s_{N+1} \equiv s_1$. For concreteness we take $s_i \in \{-1, +1\}$.[1]

If $0 < h < v$ then a descent algorithm cannot reliably find the minimum of H unless its moveset includes flips of at least $2v/h$ adjacent variables. To see this, consider global minimum configuration $s_i = 1, \forall i$ with energy $H(0, N) = (-h - v)N$. Now the configuration $s_i = -1, \forall i$ with higher energy $H(N, 0) = (h - v)N$ is a local minimum since flipping k *consecutive* variables to 1 raises the energy to $H(N-k, k) = H(N, 0) - 2kh + 4v > H(N, 0)$ for $k < 2v/h$. Flipping non-consecutive variables is higher still. There is a "barrier" between the "all-down state" and the global minimum "all-up state". Evidently a successful descent algorithm requires ingenuity to construct its moveset. MFA provides an alternative that requires no such ingenuity.

A mean field approximation of the associated Gibbs density is obtained by taking $H_0 = -\sum_i x_i s_i$ with adjustable mean field parameters $x = \{x_i\}_{i=1}^{i=N}$. The sum over configurations becomes $\sum_s \equiv \sum_{s_1=\pm 1} \sum_{s_2=\pm 1} \cdots \sum_{s_N=\pm 1}$ so that $Z_0 = \prod_i 2\cosh x_i/T$, from which $F_0 = -T \sum \ln 2\cosh x_i/T$. The average

[1] Any other kind of random binary variable in a Hamiltonian can be algebraically transformed to the range $\{-1, +1\}$, although special consideration is required if a factor of s_i^2 occurs anywhere.

value of $\langle s_i \rangle$ under H_0 is related to x_i by the function m

$$m(x_i) \equiv \tanh(x_i/T) \tag{6}$$

which tends toward the algebraic sign of x_i for small enough T so that the original binary character of the problem is recovered from MFA at $T = 0$. Now $\langle H_0 \rangle = -\sum_i x_i m(x_i)$ and

$$\langle H \rangle = -h \sum_i m(x_i) - v \sum_i m(x_i) m(x_{i+1}). \tag{7}$$

Combining these we obtain a bound on the free energy $F \leq W$

$$W = \sum_i (-T \ln 2 \cosh(x_i/T) - hm(x_i) - vm(x_i)m(x_{i+1}) + x_i m(x_i)) \tag{8}$$

which can be minimized with respect to the mean field x. The resulting self-consistency condition on x

$$x_i = h + vm(x_{i-1}) + vm(x_{i+1}) \tag{9}$$

can be converted to an equivalent condition for the expectations $m_i = m(x_i)$:

$$m_i = \tanh\left(\frac{h + vm_{i-1} + vm_{i+1})}{T}\right). \tag{10}$$

The minimum of W tells us about the thermal equilibrium at T of a system described by H, but to find the minimum of H we must in general anneal. For our shift invariant choice of h and v, the low W states have shift invariant $m_i = m, \forall i$ also. Equation 10 becomes $m = \tanh((h + 2vm)/T)$ which is graphically solved in Figure 1 for $h = v = 1$ at several values of T. Figure 1 shows that Equation 10 has one solution for high T but three solutions below a certain "critical" T. It can be shown that the middle solution is a maximum of W and the two extreme solutions are minima. Here the rightmost solution determines the global minimum as T falls. On way to describe this is that MFA varies the non-convexity of the optimization problem as a function of T.

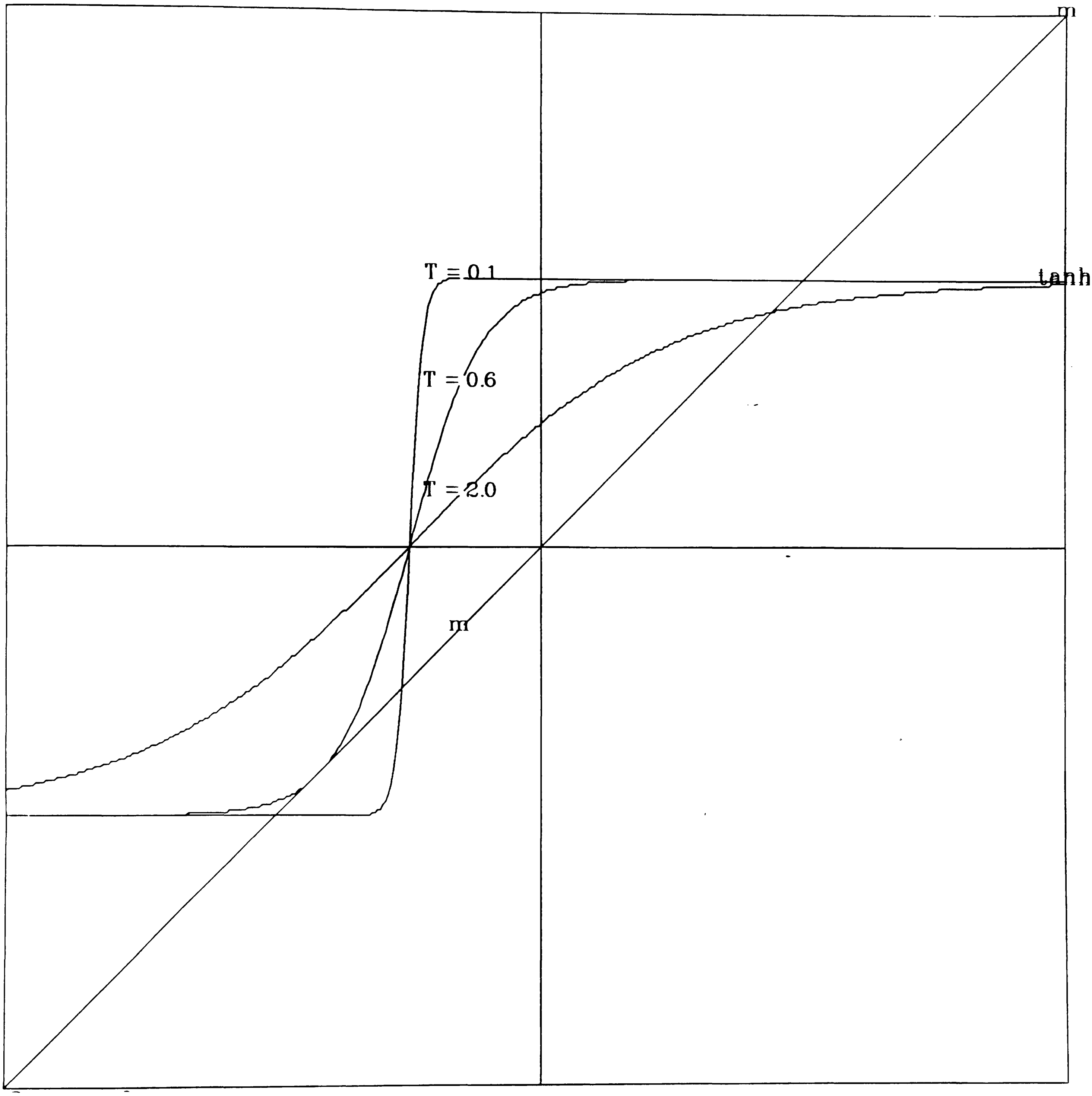

Figure 1: Self consistency Equation 10. Graphical solution for $m = \tanh((h + 2vm)/T)$. The straight line is $y = x$. The curves are $y = \tanh((h + 2vx)/T)$. Intersections between the straight lines and curves indicate extrema of W. At high T there is only one solution. At low T, the problem becomes non-convex. MFA tracks the first intersection to the lobal minimum.

3 A real-valued problem

The objective or Hamiltonian

$$H[u] = 1/2 \sum_i (u_i - d_i)^2 - \lambda \sum_i \delta(u_i - u_{i+1}) \tag{11}$$

is minimized by a piecewise constant approximation to d, *i.e.*, when real variables u_i can be chosen to resemble data d_i and to frequently coincide with neighbors $u_{i\pm 1}$. For this problem the simplest useful choice is $H_0 = \sum_i 1/2(u_i - \mu_i)^2$ with real means μ_i. The average in Peierls' inequality is weighted by the Gaussian density $Z_0^{-1} \exp(-H_0/T)$ now extends over all real configurations. Z_0 involves $\int_\infty^\infty du_1 \int_\infty^\infty du_2 \ldots \int_\infty^\infty du_N$ but it factors so that the normalization is $Z_0 = \prod_i \sqrt{2\pi T}$. The logarithm of a product is a sum of logarithms so that $F_0 = -T \sum_i \ln \sqrt{2\pi T}$. The average H_0 is a sum of second moments $\langle H_0 \rangle = \sum_i T/2 = NT/2$. For this H_0, both F_0 and $\langle H_0 \rangle$ are independent of m. All the m dependence appears in $\langle H \rangle$ which is the sum of a data term $\sum_i 1/2 \langle (u_i - d_i)^2 \rangle = NT/2$ and an interaction that involves the average of a Dirac delta $\langle \delta(u_i - u_{u+i}) \rangle$

$$\int \frac{du_i du_{i+1}}{2\pi T} \exp(-(u_i - \mu_i)^2 - (u_{i+1} - \mu_{i+1})^2) \delta(u_i - u_{i+1}) \tag{12}$$

which can be integrated once with the Dirac delta

$$\int \frac{du_i}{2\pi T} \exp(-(u_i - \mu_i)^2/T - (u_i - \mu_{i+1})^2/T) \tag{13}$$

and then evaluated with a table of integrals to obtain

$$\langle \delta(u_i - u_{i+1}) \rangle = \frac{1}{\sqrt{\pi T}} \exp(-(\mu_i - \mu_{i+1})^2/4T). \tag{14}$$

Combining these results

$$W = constant + \sum_i \left[\frac{(\mu_i - d_i)^2}{2} - \frac{\lambda}{\sqrt{\pi T}} \exp\left(-\frac{(\mu_i - \mu_{i+1})^2}{4T} \right)) \right] \tag{15}$$

where the constant is independent of μ and can be dropped. This W has the same form as the original H except that the Dirac delta has been replaced by a Gaussian which is plotted in Figure 2 for several T. Note that the

Gaussian tends to a Dirac delta as T vanishes. Figure 2 should be compared with Figure 3 where the non-convexity of the clipped parabola is also reduced by increasing T. In both cases MFA prescribes that the high-T solution be followed as T is gradually reduced. MFA replaces a non-convex problem with a family of problems of graduated non-convexity.

Two-dimensional versions of this formulation has been used to remove noise from corrupted observations of real-valued images known to be piecewise-constant[8]. MFA has been used to identify the most effective form for the delta function interaction in two dimensions[7]. Piecewise-linear restorations are similarly obtained by replacing the first difference in the argument of the delta function by the appropriate second difference[2].

4 A mixed problem

A one-dimensional version of Blake and Zisserman's weak membrane energy function[4] is (the weak string function)

$$H[u,l] = \sum_{i=1}^{i=N} \left[(u_i - d_i)^2 + \lambda^2 (u_i - u_{i+1})^2 (1 - l_i) + \alpha l_i \right], \tag{16}$$

for binary line field $l = \{l_i\}_{i=1}^{i=N}$ and $l_i \in \{0,1\}$, real intensities $u = \{u_i\}_{i=1}^{i=N}$ and $u_i \in (-\infty, \infty)$ and data $d = \{d_i\}_{i=1}^{i=N}$. For simplicity take $u_{N+1} \equiv u_1$. The parameter λ is an elastic constant and α controls when the elasticity saturates.

It is possible to treat both u and l as random fields and apply mean field theory to both, but Blake and Zisserman are concerned with non-convexities due to discrete l, so restrict H_0 to the energy of l given the mean field x

$$H_0 = -\sum_i x_i l_i. \tag{17}$$

Working toward Peierls' inequality, we will compute averages with the density $Z_0^{-1} \exp(-H_0/T)$. The normalization involves the sum over all configurations $\sum_{l_1=0,1} \sum_{l_2=0,1} \cdots \sum_{l_N=0,1}$. But the sum of products factors into a product of sums to give $Z_0 = \prod_i (1 + \exp(x_i/T))$. From this $F_0 = -T \sum_i \ln(1 + \exp(x_i/T))$. The expectation of the random variable l_i are related to the mean field parameter x_i as $\langle l_i \rangle = \sigma(x_i)$ where

$$\sigma(x_i) = \frac{0 + e^{x_i/T}}{1 + e^{x_i/T}} = \frac{1}{1 + e^{-x_i/T}}. \tag{18}$$

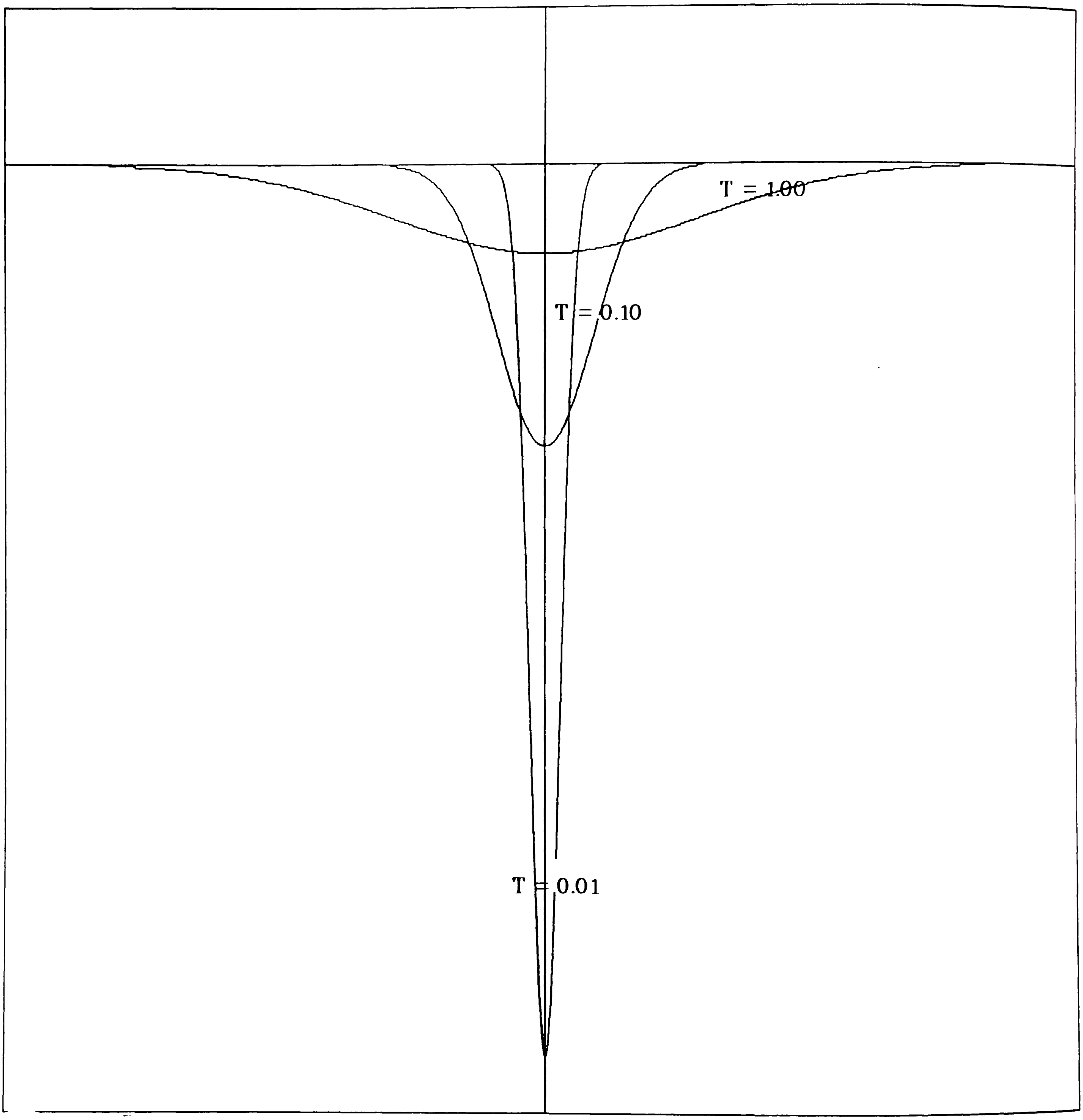

Figure 2: The interaction potential for Equation 15 plotted for several T. Note the gradual reduction in non-convexity as T is increased.

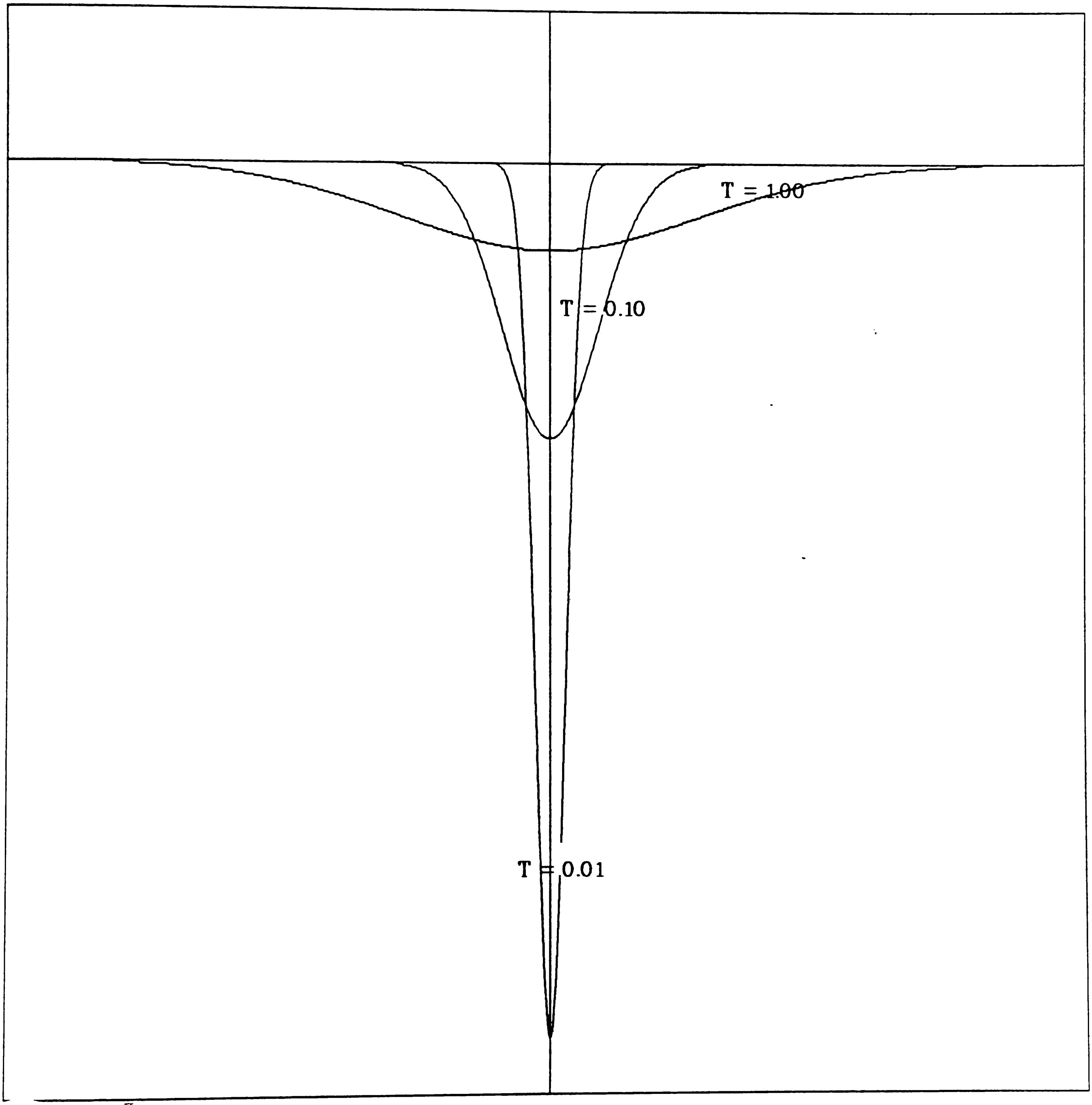

Figure 2: The interaction potential for Equation 15 plotted for several T. Note the gradual reduction in non-convexity as T is increased.

In terms of σ

$$\langle H_0 \rangle = -\sum_i x_i \sigma(x_i). \tag{19}$$

and

$$\langle H \rangle = \sum_{i=1}^{i=N} \left[(u_i - d_i)^2 + \lambda^2 (u_i - u_{i+1})^2 (1 - \sigma(x_i)) + \alpha \sigma(x_i) \right]. \tag{20}$$

Combining these, the Weiss free energy W is

$$\sum_i \left[-T \ln(1 + e^{x_i/T}) + (u_i - d_i)^2 + \lambda^2 (u_i - u_{i+1})^2 (1 - \sigma(x_i)) + \alpha \sigma(x_i) + x_i \sigma(x_i) \right]. \tag{21}$$

The optimal value of x is determined by solving the system

$$\frac{\partial W}{\partial x_i} = \left(-\lambda^2 (u_i - u_{i+1})^2 + \alpha + x_i \right) \frac{\partial \sigma(x_i)}{\partial x_i} = 0, \forall i \tag{22}$$

which occurs when

$$x_i = \lambda^2 (u_i - u_{i+1})^2 - \alpha \tag{23}$$

which depends on the u field. Substituting Equation 23 into Equation 21 yields

$$\min_x W(u, x) = \sum_i \left[(u_i - d_i)^2 - T \ln \left(1 + \exp \frac{\lambda^2 (u_i - u_{i+1})^2 - \alpha}{T} \right) + \lambda^2 (u_i - u_{i+1})^2 \right], \tag{24}$$

which is then to be minimized with respect to u (here by simple gradient descent) to determine the mean field approximation to thermal equilibrium at T. MFA prescribes that the minimum $\min_{ux} W(u,x) = \min_u \min_x W(u,x)$ be annealed to find the minimum of H, but this is essentially GNC as was first reported by Geiger and Girosi who called the last two terms of Equation 24 the "effective potential". In Figure 3, this effective potential is plotted against $u_i - u_{i+1}$ for several values of T. Note that for $T \leq .2$ this plot is qualitatively identical to the GNC heuristic for smoothing of the weak membrane potential. Thus at low T, the MFA $\min_x W(u,x)$ approximately reproduces Blake and Zissermans piecewise quadratic heuristic function for u. At higher T Figure 3 also shows that the MFA effective potential is deeper than the GNC. In the limit of high T the MFA effective potential approaches $-T \ln 2 + (u_i - u_{i+1})^2$.

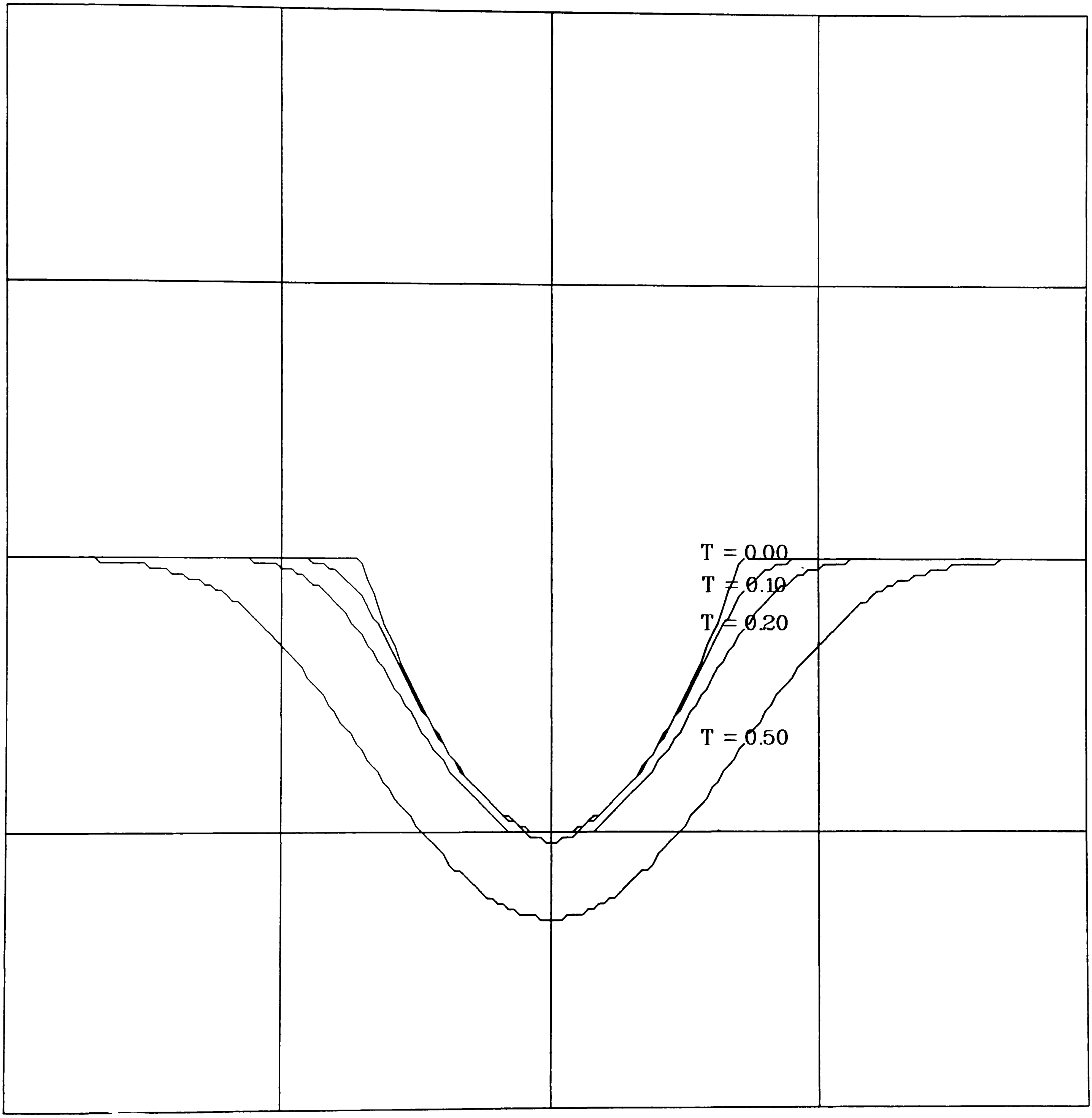

Figure 3: A plot of $-T \ln\left(1 + \exp\frac{\lambda^2(u_i - u_{i+1})^2 - \alpha}{T}\right) + \lambda^2(u_i - u_{i+1})^2$ the last two terms of Equation 23 against the difference $u_i - u_{i+1}$ for several values of T.

The mean field approximation itself does not explain the superiority of MFA piecewise-constant restorations over GNC weak-membrane restorations. MFA applied to the weak membrane approximates Blake and Zissermans GNC algorithm. MFA applied to an objective with a delta function interaction produces our previous MFA algorithms for piecewise constant restoration. We conclude that MFA produces better piecewise constant restorations than GNC because a delta function interaction is a better model of piecewise constant surfaces than is the weak membrane. The weak membrane of GNC models the scene with a piecewise smooth objective which is less appropriate for scenes that are really piecewise constant.

References

[1] G. L. Bilbro, T. K. Miller, W. E. Snyder and D. E. Van den Bout, M. W. White, and R. C. Mann. Optimization by the Mean Field Approximation. In *Advances in Neural Network Information Processing Systems 1*, pages 91–98. Morgan-Kauffman, San Mateo, 1989. Reprinted from 1988 NIPS, Denver, CO.

[2] G. L. Bilbro and W. E. Snyder. Mean field annealing: An application to image noise removal. *Journal of Neural Network Computing*, 1990.

[3] G. L. Bilbro, W. E. Snyder, and R. C. Mann. The mean field minimizes relative entropy. *Journal of the Optical Society of America A*, 1991. To appear.

[4] A. Blake and A. Zisserman. *Visual Reconstruction.* The MIT Press, Cambridge, Mass., 1987.

[5] D. E. Van den Bout and T. K. Miller. Graph partitioning using annealed neural networks. *IEEE Trans. Neural Networks*, 1(2):192–203, 1990.

[6] D. Geiger and F. Girosi. Parallel and deterministic algorithms for mrfs: surface reconstructin and integration. Technical Report A. I. Memo No. 1114, Massachusetts Institute of Technology, 1989.

[7] H. P. Hiriyannaiah. *Signal reconstruction using mean field annealing.* PhD thesis, North Carolina State University, 1990.

[8] H. P. Hiriyannaiah, G. L. Bilbro, W. E. Snyder, and R. C. Mann. Restoration of piecewise constant images via mean field annealing. *Journal of the Optical Society of America A*, pages 1901–1912, December 1989.

[9] C. M. Soukoulis, K. Levin, and G. S. Grest. Irreversibility and metastability in spin-glasses. i ising model. *Phys. Rev. B*, 28(3):1495–1509, August 1983.

[10] P. J. M. van Laarhoven and E. H. L. Aarts. *Simulated Annealing: Theory and Applications.* D. Reidel, Norwell, Mass, 1988.