GRAPHICAL MODELS AND IMAGE PROCESSING
Vol. 58, No. 1, January, pp. 18–37, 1996
ARTICLE NO. 0002

A Hierarchical Markov Random Field Model and Multitemperature Annealing for Parallel Image Classification*

ZOLTAN KATO, MARC BERTHOD, AND JOSIANE ZERUBIA

INRIA, 2004 Route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex, France

Received July 27, 1993; in revised form August 4, 1995; accepted September 19, 1995

In this paper, we are interested in massively parallel multiscale relaxation algorithms applied to image classification. It is well known that multigrid methods can improve significantly the convergence rate and the quality of the final results of iterative relaxation techniques. First, we present a classical multiscale model which consists of a label pyramid and a whole observation field. The potential functions of coarser grids are derived by simple computations. The optimization problem is first solved at the higher scale by a parallel relaxation algorithm; then the next lower scale is initialized by a projection of the result. Second, we propose a hierarchical Markov random field model based on this classical model. We introduce new interactions between neighbor levels in the pyramid. It can also be seen as a way to incorporate cliques with far apart sites for a reasonable price. This model results in a relaxation algorithm with a new annealing scheme: the multitemperature annealing (MTA) scheme, which consists of associating higher temperatures to higher levels, in order to be less sensitive to local minima at coarser grids. The convergence to the global optimum is proved by a generalization of the annealing theorem of S. Geman and D. Geman (*IEEE Trans. Pattern Anal. Mach. Intell.* 6, 1984, 721–741). © 1996 Academic Press, Inc.

1. INTRODUCTION

Markov random fields (MRF) have become more and more popular during the past few years in image processing [1, 5, 8, 9, 12, 27]. A good reason for this is that MRF require less a priori information on the world model. On the other hand, the local behavior of MRF allows for the development of highly parallel algorithms in combinatorial optimization problems.

In this paper, we are interested in massively parallel multiscale relaxation algorithms applied to image classification [6, 7, 14, 15]. It is well known that multigrid methods can improve significantly the convergence rate and the quality of the final results of iterative relaxation techniques.

There are many approaches in multigrid image segmentation. A well known approach is the renormalization group algorithm which is based on renormalization group ideas from statistical physics. This technique has been adapted by Gidas [13] to image processing. The main advantage of the method is that it provides a mechanism for relating the processing at different scales with one another. This mechanism is a nonlinear transformation—called the *renormalization group* (RG) *transformation*. The coarser grids and their Hamiltonians are well defined; they are deduced from the original image. The major difficulty is the computation of the energy functions at coarser grids. Usually, this computation cannot be done explicitly; one must approximate them [10, 25]. Another drawback is the loss of Markovianity at coarser grids [26] since the coarser energy functions obtained by the RG transformation cannot be decomposed as a sum of clique-potentials. In [13], the Hamiltonians are approximated by a sum of clique-potentials, and hence one can use classical relaxation algorithms to minimize the energy at coarser grids. Unfortunately, such approximations are available only for certain simple models, mainly in image restoration [13]. Another interesting model has been proposed by Bouman and Shapiro [7]. This model consists of a label pyramid where each level is causally dependent on the coarser layer above it. The model results in a new optimization criterium called *sequential MAP* estimate. This model yields to a noniterative segmentation algorithm and direct methods of parameter estimation.

The basis of our approach is a consistent multiscale MRF model originally proposed by Heitz *et al.* in [14, 15] for motion analysis. Related models can also be found in [6] for texture segmentation and in [17] for image reconstruction. This model consists of a label pyramid and a whole observation field. The original energy function can be decomposed as a sum of potential functions which are defined

* This work has been partially funded by CNES (French Space Agency), AFIRST, and GdR TdSI (DRED/CNRS). We thank the second reviewer of the paper for pointing out a lack in the Multitemperature Annealing Theorem.

on neighbor blocks and only depend on the labels associated with these blocks and on the observation field. Using this decomposition, the parameters of coarser grids can be computed very easily. This model results in a multigrid relaxation scheme which replaces the original optimization problem by a sequence of more tractable problems. Using a top down strategy in the label pyramid, the optimization problem is first solved at a higher level; then the lower grid is initialized with the previous result by a simple projection. This algorithm is very efficient in the case of deterministic relaxation (for instance, ICM [3, 18]) which gets stuck in a local minimum near the starting configuration. In the case of stochastic relaxation (for instance, simulated annealing [11, 23, 24]), which is far less dependent on the initial configuration, the results are only slightly better, but the method is still interesting with respect to computer time, especially on a sequential machine. We give a general description of this model and the relaxation scheme associated with it in Section 2.

Then we propose a new hierarchical MRF model defined on the whole label pyramid (Section 3). In this model, we have introduced a new interaction scheme between neighboring levels in the pyramid, yielding a better communication between the grids. It can also be seen as a way to incorporate cliques with far apart sites for a reasonable price. This model gives a relaxation algorithm with a new annealing scheme which can be run in parallel on the entire pyramid. The basic idea of this annealing scheme, which we propose to call multitemperature annealing (MTA), is the following: to the higher levels, we associate higher temperatures which enable the algorithm to be less sensitive to local minima. However at a finer resolution, the relaxation is performed at a lower temperature. The complete convergence study of the multitemperature annealing schedule can be found in Section 4. Our annealing theorem is a generalization of the well-known theorem of Geman and Geman [11] and the proof can be found in the Appendix.

Finally, image segmentation experiments are shown in Section 5 with the Gibbs sampler [11] and the ICM [3, 18] using the three models for each algorithm (monogrid, multiscale, and hierarchical). These methods have been implemented in parallel on Connection Machine CM200 [16].

2. MULTISCALE MRF MODELS

Herein, we are interested in the following general problem: we are given a set of sites $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$ and a set of image data $\mathcal{F} = \{f_s\}_{s\in\mathcal{S}}$. Each of these sites may belong to any one of L classes (or equivalently take any label from $\Lambda = \{1, 2, \ldots, L\}$). A global discrete labeling ω assigns one label ω_s ($\omega_s \in \Lambda$) to each site s in $\mathcal{S}$. $(\omega, \mathcal{F})$ is an MRF with respect to a chosen neighborhood system $\mathcal{G} = \{\mathcal{G}_s\}_{s\in\mathcal{S}}$.

Let us consider now a monogrid supervised image segmentation model [21, 22] and suppose that the observations consist of gray levels. A very general problem is to find the labeling $\hat{\omega}$ which maximizes $P(\omega \mid \mathcal{F})$. Bayes theorem tells us that $P(\omega \mid \mathcal{F}) = (1/P(\mathcal{F}))\, P(\mathcal{F} \mid \omega)P(\omega)$. Actually $P(\mathcal{F})$ does not depend on the labeling ω and we have the assumption that $P(\mathcal{F} \mid \omega) = \prod_{s\in\mathcal{S}} P(f_s \mid \omega_s)$. It is then easy to see that, under some independence assumption [4], the global labeling which we are trying to find is given by

$$\hat{\omega} = \max_{\omega\in\Omega} \prod_{s\in\mathcal{S}} P(f_s \mid \omega_s) \prod_{C\in\mathcal{C}} \exp(-V_C(\omega_C)). \tag{1}$$

It is obvious from this expression that the a posteriori probability also derives from an MRF. The energies of cliques of order 1 directly reflect the probabilistic modeling of labels without context, which would be used for labeling the pixels independently. Let us assume that $P(f_s \mid \omega_s)$ is Gaussian, the class $\lambda \in \Lambda$ is represented by its mean value μ_λ, and its deviation is represented by σ_λ. We get the energy function

$$U = \sum_{s\in\mathcal{S}} \left(\log(\sqrt{2\pi}\sigma_{\omega_s}) + \frac{(f_s - \mu_{\omega_s})^2}{2\sigma^2_{\omega_s}} \right) + \sum_{C\in\mathcal{C}} V_2(\omega_C), \tag{2}$$

where

$$V_2(\omega_C) = V_{\{s,r\}}(\omega_s, \omega_r) = \begin{cases} -\beta & \text{if } \omega_s = \omega_r \\ +\beta & \text{if } \omega_s \neq \omega_r \end{cases} \tag{3}$$

with $\beta \geq 0$.

The initial problem is reduced to a combinatorial optimization problem, namely to the minimization of a nonconvex energy function. Several approaches have been proposed to solve this task, such as simulated annealing (SA) [11, 23, 24], ICM [3, 18], and modified metropolis dynamics (MMD) [22]. Multigrid schemes have also been proved to be very efficient for energy minimization [7]. Here, we briefly describe a classical multiscale model extensively studied by Heitz *et al.* in [14, 15], which was the basis for our hierarchical MRF model.

2.1. The Classical Multiscale Model

In the following, we will focus on a MRF defined over a first-order neighborhood system with an energy function given by

$$U(\omega, \mathcal{F}) = U_1(\omega, \mathcal{F}) + U_2(\omega), \tag{4}$$

where U_1 (resp. U_2) denotes the energy of the first-order (resp. second-order) cliques.

To generate a multigrid MRF model, let us divide the

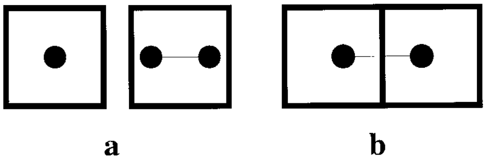


FIG. 1. The two subsets of $\mathscr{C}$ in the case of a first-order neighborhood system. (a) $\mathscr{C}^i_k$; (b) $\mathscr{C}^i_{k,l}$.

initial grid into blocks of $n \times n$, typically 16 (4×4) neighbor pixels. We consider that the label assigned to the pixels of a block is constant over all the pixels of the block. These configurations will describe the MRF at scale 1. Scale i is defined similarly by considering labels which are constant over blocks of size $n^i \times n^i$.

Let $\mathscr{B}^i = \{b^i_1, \ldots, b^i_{N_i}\}$ ($N_i = N/n^{2i}$) denote the set of blocks and Ω_i the configuration space at scale i ($\Omega_i \subset \Omega_{i-1} \subset \cdots \subset \Omega_0 = \Omega$). The label associated with block b^i_k is denoted by ω^i_k. We can define the same neighborhood structure on $\mathscr{B}^i$ as on $\mathscr{S}$:

$$b^i_k \text{ and } b^i_l \text{ are neighbors}$$
$$\Leftrightarrow \begin{cases} b^i_k \equiv b^i_l \text{ or} \\ \exists C \in \mathscr{C} \mid C \cap b^i_k \neq \emptyset \text{ and } C \cap b^i_l \neq \emptyset. \end{cases} \tag{5}$$

Now, let us partition the original set $\mathscr{C}$ into two disjoint subsets $\mathscr{C}^i_k$ and $\mathscr{C}^i_{k,l}$:

1. Cliques which are included in b^i_k (see Fig. 1a):

$$\mathscr{C}^i_k = \{C \in \mathscr{C} \mid C \subset b^i_k\}. \tag{6}$$

2. Cliques which sit astride two neighboring blocks $\{b^i_k, b^i_l\}$ (see Fig. 1b):

$$\begin{aligned} \mathscr{C}^i_{k,l} = \{C \in \mathscr{C} \mid C \subset (b^i_k \cup b^i_l) \text{ and} \\ C \cap b^i_k \neq \emptyset \text{ and } C \cap b^i_l \neq \emptyset\}. \end{aligned} \tag{7}$$

It is obvious from this partition that our energy function (see Eq. (4)) can be decomposed as

$$\begin{aligned} U_1(\omega, \mathscr{F}) &= \sum_{s \in \mathscr{S}} V_1(\omega_s, f_s) \\ &= \sum_{b^i_k \in \mathscr{B}^i} \underbrace{\sum_{s \in b^i_k} V_1(\omega_s, f_s)}_{V^{\mathscr{B}^i}_1(\omega^i_k, \mathscr{F})} = \sum_{b^i_k \in \mathscr{B}^i} V^{\mathscr{B}^i}_1(\omega^i_k, \mathscr{F}) \end{aligned} \tag{8}$$

and

$$\begin{aligned} U_2(\omega) &= \sum_{C \in \mathscr{C}} V_2(\omega_c) \\ &= \sum_{b^i_k \in \mathscr{B}^i} \underbrace{\sum_{C \in \mathscr{C}^i_k} V_2(\omega_c)}_{V^{\mathscr{B}^i}_k(\omega^i_k)} + \sum_{\{b_k, b_l\} neighbors} \underbrace{\sum_{C \in \mathscr{C}^i_{k,l}} V_2(\omega_c)}_{V^{\mathscr{B}^i}_{k,l}(\omega^i_k,\, \omega^i_l)} \\ &= \sum_{b^i_k \in \mathscr{B}^i} V^{\mathscr{B}^i}_k(\omega^i_k) + \sum_{\{b_k, b_l\} neighbors} V^{\mathscr{B}^i}_{k,l}(\omega^i_k, \omega^i_l). \end{aligned} \tag{9}$$

Now, we can define our pyramid (see Fig. 2) where level i contains the coarse grid $\mathscr{S}^i$ which is isomorphic to the scale $\mathscr{B}^i$. The coarse grid has a reduced configuration space $\Xi^i = \Lambda^{N_i}$. The isomorphism Φ^i from $\mathscr{S}^i$ in $\mathscr{B}^i$ is just a projection of the coarse label field to the fine grid $\mathscr{S}^0 = \mathscr{S}$:

$$\begin{aligned} \Phi^i \colon \Xi^i &\rightarrow \Omega_i \\ \xi^i &\mapsto \omega = \Phi^i(\xi^i). \end{aligned} \tag{10}$$

The model on the grids $\mathscr{S}^i$ ($i = 0, \ldots, M$) defines a set of consistent multiscale MRF models, whose energy functions derived from Eqs. (8) and (9)

$$\begin{aligned} U^i(\xi^i, \mathscr{F}) &= U^i_1(\xi^i, \mathscr{F}) + U^i_2(\xi^i) \\ &= U_1(\Phi^i(\omega_i), \mathscr{F}) + U_2(\Phi^i(\omega_i)) \; i = 0, \ldots, M \end{aligned} \tag{11}$$

where

$$\begin{aligned} U^i_1(\xi^i, \mathscr{F}) &= \sum_{k \in \mathscr{S}^i} (V^{\mathscr{B}^i}_1(\omega^i_k, \mathscr{F}) + V^{\mathscr{B}^i}_k(\omega^i_k)) \\ &= \sum_{k \in \mathscr{S}^i} V^i_1(\omega^i_k, \mathscr{F}) \end{aligned} \tag{12}$$

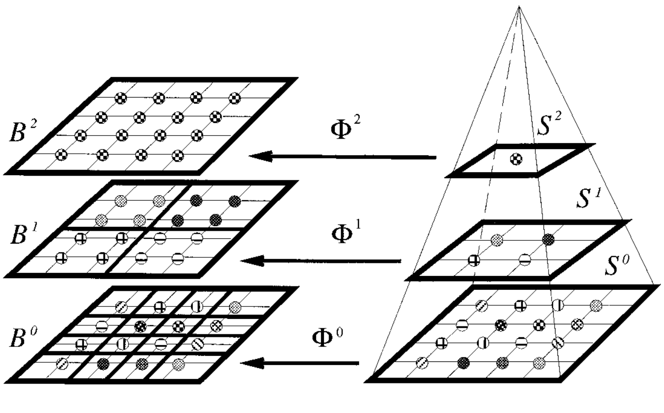


FIG. 2. The isomorphism Φ^i between $\mathscr{B}^i$ and $\mathscr{S}^i$.

and

$$U_2^i(\xi^i) = \sum_{\{k,l\} neighbors} V_{k,l}^{\mathscr{B}^i}(\omega_k^i, \omega_l^i) = \sum_{C^i \in \mathscr{C}^i} V_2^i(\omega_C^i), \quad (13)$$

where C^i is a second-order clique corresponding to the definition in Eq. (5) and $\mathscr{C}^i$ is the set of cliques on the grid i.

The relaxation scheme on this pyramid is very simple. Instead of the original optimization problem, we have a sequence of problems to solve:

$$\hat{\omega}^i = \arg\min_{\xi^i \in \Xi^i} U^i(\xi^i, \mathscr{F}), i = M, \ldots, 0. \quad (14)$$

Using a top-down strategy in the pyramid, we solve the problem first at a higher level i; then the level $i - 1$ is initialized by $(\Phi^{i-1})^{-1} \circ \Phi^i(\hat{\omega}^i)$, where $\hat{\omega}^i$ is obtained at the convergence of a relaxation algorithm at level i. Before explaining in detail the relaxation scheme, let us derive the equations of a multiscale image segmentation model using the results reported in Eqs. (2) and (3) [19, 20]:

$$U_1^i(\xi^i, \mathscr{F}) = \sum_{s^i \in \mathscr{S}^i} V_1^i(\xi_{s^i}^i, \mathscr{F}),$$

where

$$\begin{aligned} V_1^i(\xi_{s^i}^i, \mathscr{F}) &= \sum_{s \in b_{s^i}^i} V_1(\omega_s, f_s) + \sum_{C \in \mathscr{C}_{s^i}^i} V_2(\omega_C) \\ &= \sum_{s \in b_{s^i}^i} \left(\log(\sqrt{2\pi}\sigma_{\omega_s}) + \frac{(f_s - \mu_{\omega_s})^2}{2\sigma_{\omega_s}^2} \right) - p^i\beta \end{aligned} \quad (15)$$

and

$$U_2^i(\xi^i) = \sum_{C^i = \{r^i, s^i\} \in \mathscr{C}^i} V_2^i(\xi_{C^i}^i),$$

where

$$V_2^i(\xi_{C^i}^i) = \sum_{\{r,s\} \in \mathscr{D}_{C^i}} V_2(\omega_r, \omega_s) = \begin{cases} -q^i\beta & \text{if } \omega_r = \omega_s \\ +q^i\beta & \text{if } \omega_r \neq \omega_s. \end{cases} \quad (16)$$

The values of p^i and q^i depend on the chosen block size and the neighborhood structure, p^i is the number of cliques included in the same block at scale $\mathscr{B}^i$ and q^i is the number of cliques between two neighboring blocks at scale $\mathscr{B}^i$. Considering blocks of $n \times n$ and a first-order neighborhood system, we get

$$p^i = 2n^i(n^i - 1) \quad (17)$$

$$q^i = n^i. \quad (18)$$

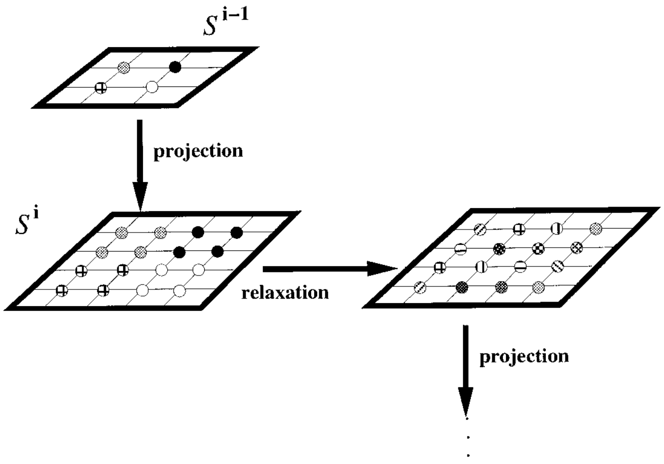


FIG. 3. The multiscale relaxation scheme.

2.2. Relaxation Scheme

Basically, this is the same procedure as in the monogrid case. The only difference is that we have more functions to minimize which are less complex than the original one. The algorithm is the following (see Fig. 3): instead of minimizing the original energy function U, we tackle the sequence of problems U^i ($M \geq i \geq 0$) using a top-down strategy in the pyramid. First, we solve the problem at a higher level i using a parallel relaxation scheme; then the level $i - 1$ is initialized by $(\Phi^{i-1})^{-1} \circ \Phi^i(\hat{\xi}^i)$ which is just a projection of $\hat{\xi}^i$ on the finer grid $\mathscr{S}^{i-1}$ ($\hat{\xi}^i$ is the solution at level i).

The advantages of this algorithm are clear: each $\hat{\xi}^i$ gives a more or less good estimate of the final result. The estimate is better as i goes down to 0. On the other hand, for the higher values of i, the corresponding problem is simpler since the state space has only a few elements.

The scheme is particularly well adapted to the deterministic relaxation methods which are more sensitive to the initial configuration than the stochastic ones. In our experiments (see Section 5), the final result is improved compared to the monogrid version of the same algorithm. However, for the stochastic ones, the final result is only slightly improved since these methods are independent of the initial configuration.

Another important measure of the efficiency is the speed of convergence. On a sequential machine, the proposed scheme exhibits fast convergence properties. However, on a SIMD machine, the speed depends mainly on the virtual processor ratio (VPR = number of the virtual processors per physical processor). This means that the monogrid scheme may be faster on such a machine, considering the (very simple) parallelization described above, because the multiscale scheme demands usually more iterations (the relaxation algorithm must converge at each level and there is a minimal number of iterations necessary for the convergence). In our experiments, the monogrid scheme was always faster than this scheme on a Connection Machine CM200 (see Section 5).

We note that in [15] another parallelization scheme has been proposed which consists of generating configurations in parallel, using different temperatures at different levels, with periodic interactions between them. The interaction introduces a transfer, at every n iterations, of a small block of labels to the next finer level. The block is accepted, if the energy of the new block is lower (deterministic rule). We also implemented a finer version of this scheme. In our approach, each site at each iteration transfers its state to the next lower level. At the lower scale, this information is taken into account as the state of an additional neighbor site. The transition is then governed by the Gibbs sampler or any other method, taking into account this external information (probabilistic rule).

The problem with both algorithms is that, to our knowledge, the convergence of such an algorithm has not been proved. Looking for a better parallelization scheme with a theoretical background may be a future work.

3. THE HIERARCHICAL MODEL

In this section, we propose a new hierarchical MRF model. The basic idea is to find a better way of communication between the levels than the initialization used for the multiscale model. Our approach consists in introducing new interactions between two neighbor grids[1] in the pyramid. This scheme permits also the parallelization of the relaxation algorithm on the whole pyramid. First, we give a general description of the model; then we study a special case with a first-order neighborhood system.

3.1. General Description

We consider here the label pyramid and the whole observation field defined in the previous section. Let $\overline{\mathscr{S}} = \{\bar{s}_1, \ldots, \bar{s}_{\overline{N}}\}$ denote the sites of this pyramid. Obviously,

$$\overline{\mathscr{S}} = \bigcup_{i=0}^{M} \mathscr{S}^i \qquad \overline{N} = \sum_{i=0}^{M} N_i. \tag{19}$$

$\overline{\Omega}$ denotes the configuration space of the pyramid:

$$\overline{\Omega} = \Xi^0 \times \Xi^1 \times \cdots \times \Xi^M = \{\overline{\omega} \mid \overline{\omega} = (\xi^0, \xi^1, \ldots, \xi^M)\}. \tag{20}$$

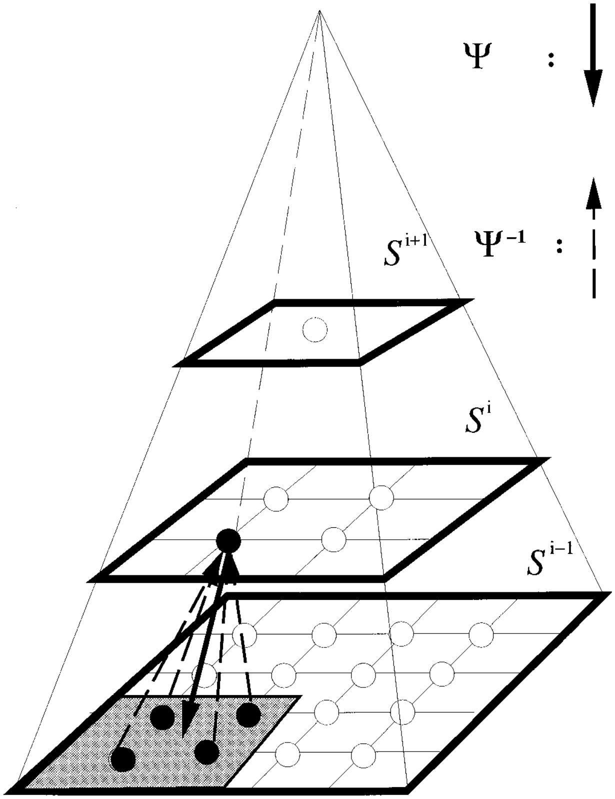


FIG. 4. The functions Ψ and Ψ^{-1}.

Let us define the following function Ψ between two neighbor levels, which assigns to a site of any level the corresponding block of sites at the level below it (that is, its descendants). Ψ^{-1} assigns its ancestor to a site (see Fig. 4):

$$\begin{aligned} \Psi: \quad & \mathscr{S}^i \to \mathscr{S}^{i-1} \\ \Psi(\bar{s}) &= \{\bar{r} \mid \bar{s} \in \mathscr{S}^i \Rightarrow \bar{r} \in \mathscr{S}^{i-1} \text{ and } b_{\bar{r}}^{i-1} \subset b_{\bar{s}}^{i}\}. \end{aligned} \tag{21}$$

Now we can define on these sites the neighborhood system (see Fig. 5)

$$\overline{\mathscr{G}} = \left(\bigcup_{i=0}^{M} \mathscr{G}_i\right) \cup \{\Psi^{-1}(\bar{s}) \cup \Psi(\bar{s}) \mid \bar{s} \in \overline{\mathscr{S}}\}, \tag{22}$$

where $\mathscr{G}_i$ is the neighborhood structure of the ith level, and we have the cliques

$$\overline{\mathscr{C}} = \left(\bigcup_{i=0}^{M} \mathscr{C}^i\right) \cup \mathscr{C}^*, \tag{23}$$

where $\mathscr{C}^*$ denotes the new cliques sitting astride two neighbor grids. We can easily estimate the degree of the new cliques since it depends on the block size: Each site inter-

[1] One can imagine interactions between more than two levels but these schemes are too complicated for practical use.

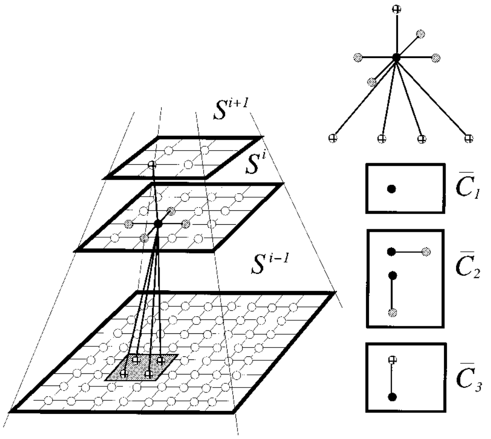


FIG. 5. The neighborhood system $\overline{\mathscr{G}}$ and the cliques $\overline{\mathscr{C}}_1$, $\overline{\mathscr{C}}_2$, and $\overline{\mathscr{C}}_3$.

acts with its ancestor (there is one) and its descendants (there are wh); thus,

$$\deg(\mathscr{C}^*) = \max_{C^* \in \mathscr{C}^*} | C^* | = wh + 2 \tag{24}$$

and

$$\deg(\overline{\mathscr{C}}) = \deg(\mathscr{C}) + \deg(\mathscr{C}^*) - 1. \tag{25}$$

Furthermore, let $\overline{\mathscr{X}}$ be a MRF over $\overline{\mathscr{G}}$ with energy function $\overline{U}$ and potentials $\{\overline{V}_{\overline{C}}\}_{\overline{C} \in \overline{\mathscr{C}}}$. The energy function is of the form

$$\begin{aligned} \overline{U}(\overline{\omega}) &= \sum_{\overline{C} \in \overline{\mathscr{C}}} \overline{V}_{\overline{C}}(\overline{\omega}) \\ &= \sum_{i=0}^{M} \sum_{\overline{C} \in \mathscr{C}^i} V^i_{\overline{C}}(\overline{\omega}) + \sum_{\overline{C} \in \mathscr{C}^*} \overline{V}_{\overline{C}}(\overline{\omega}) \\ &= \sum_{i=0}^{M} \sum_{C^i \in \mathscr{C}^i} V^i_{C^i}(\xi^i) + \sum_{C^* \in \mathscr{C}^*} \overline{V}_{C^*}(\overline{\omega}) \\ &= \sum_{i=0}^{M} U^i(\xi^i) + U^*(\overline{\omega}). \end{aligned} \tag{26}$$

It turns out from the above equation that the energy function consists of two terms. The first corresponds to the sum of the energy functions of the grids defined in the previous section and the second ($U^*(\overline{\omega})$) is the energy over the new cliques located between neighbor grids.

3.2. A Special Case

In this section, we study the model in the case of a first-order neighborhood system. We will consider herein only first- and second-order cliques. Clique potentials for the other cliques are supposed to be 0. The cliques can be partitioned into three disjoint subsets $\overline{\mathscr{C}}_1, \overline{\mathscr{C}}_2, \overline{\mathscr{C}}_3$ corresponding to first-order cliques, second-order cliques which are on the same level, and second-order cliques which sit astride two neighboring levels (see Fig. 5). Using this partition, we can derive the energy function

$$\overline{U}(\overline{\omega}, \mathscr{F}) = \overline{U}_1(\overline{\omega}, \mathscr{F}) + \overline{U}_2(\overline{\omega}) \tag{27}$$

$$\begin{aligned} \overline{U}_1(\overline{\omega}, \mathscr{F}) &= \sum_{\overline{s} \in \overline{\mathscr{S}}} \overline{V}_1(\overline{\omega}_{\overline{s}}, \mathscr{F}) \\ &= \sum_{i=0}^{M} \sum_{s^i \in \mathscr{S}^i} V^i_1(\xi^i_{s^i}, \mathscr{F}) = \sum_{i=0}^{M} U^i_1(\xi^i, \mathscr{F}) \end{aligned} \tag{28}$$

$$\begin{aligned} \overline{U}_2(\overline{\omega}) &= \sum_{C \in \overline{\mathscr{C}}_2} \overline{V}_2(\overline{\omega}_C) + \sum_{C \in \overline{\mathscr{C}}_3} \overline{V}_2(\overline{\omega}_C) \\ &= \sum_{i=0}^{M} \sum_{C \in \mathscr{C}^i} V^i_2(\xi^i_C) + \sum_{C \in \overline{\mathscr{C}}_3} \overline{V}_2(\overline{\omega}_C) \\ &= \sum_{i=0}^{M} U^i_2(\xi^i) + \sum_{C \in \overline{\mathscr{C}}_3} \overline{V}_2(\overline{\omega}_C). \end{aligned} \tag{29}$$

The equations of a hierarchical image segmentation model are (using Eqs. (28) and (29)) [19, 20]:

$$\overline{U}_1(\overline{\omega}, \mathscr{F}) = \sum_{i=0}^{M} \sum_{s^i \in \mathscr{S}^i} V^i_1(\xi^i, \mathscr{F}) \tag{30}$$

and

$$\overline{U}_2(\overline{\omega}) = \sum_{i=0}^{M} \sum_{C^i \in \mathscr{C}^i} V^i_2(\xi^i_{C^i}) + \sum_{C \in \overline{\mathscr{C}}_3} \overline{V}_2(\overline{\omega}_c), \tag{31}$$

where

$$\overline{V}_2(\overline{\omega}_c) = \overline{V}_{\{\overline{s},\overline{r}\}}(\overline{\omega}_{\overline{s}}, \overline{\omega}_{\overline{r}}) = \begin{cases} -\gamma & \text{if } \overline{\omega}_{\overline{s}} = \overline{\omega}_{\overline{r}} \\ +\gamma & \text{if } \overline{\omega}_{\overline{s}} \neq \overline{\omega}_{\overline{r}}, \end{cases} \tag{32}$$

with $\gamma \geq 0$.

In the next section, we propose a new annealing scheme for the efficient minimization of the energy function of the hierarchical model.

4. MULTITEMPERATURE ANNEALING

4.1. Parallel Relaxation Scheme

Now, let us see how the energy $\overline{U}(\overline{\omega})$ is minimized. If we use a deterministic relaxation method where the temperature parameter is kept constant during the iterations (for example, ICM [3]), then the original formulation of the algorithm does not change. The only difference is that we work on a pyramid and not on a rectangular shape as in

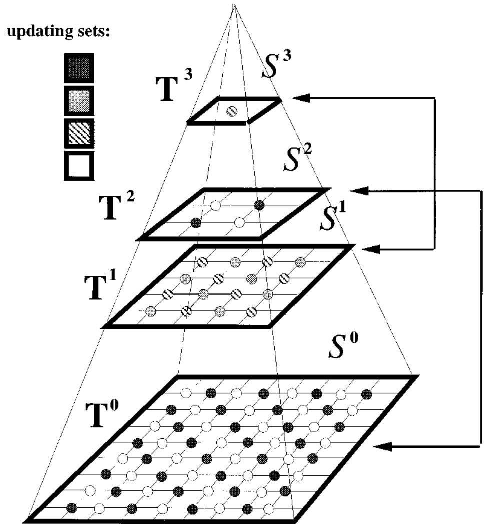


FIG. 6. Relaxation scheme on the pyramid. The levels connected by pointers are updated at the same time.

the monogrid case. We can easily parallelize this algorithm using the coding technique described by Besag in [3]: we partition the pyramid $\overline{\mathcal{S}}$ into disjoint updating sets so that pixels which belong to the same set are conditionally independent, given the data of all the other sets. This enables us to update different levels at the same time (see Fig. 6).

Let us consider in the following a relaxation algorithm where the temperature changes during the iterations. The temperature change is controlled by a function, the so-called *annealing schedule*. Such a method is, for example, the simulated annealing (Gibbs sampler [11], metropolis algorithm [23, 24]) or some deterministic scheme such as modified metropolis dynamics [21, 22]. For these algorithms, we introduce a new annealing schedule: the multi-temperature annealing (MTA). The idea is to associate different temperatures to different levels in the pyramid. For the cliques sitting between two levels, we use either the temperature of the lower level or that of the higher level (but once chosen, we always keep the same level throughout the algorithm). For the parallelization [2], we use the same coding technique as in the previous case.

We have three ways of annealing. The first two are well known [23]; they require no modification of the original algorithm, except that we work on a pyramid instead of a rectangular shape. The third is a new annealing schedule which is the most efficient with the hierarchical model:

1. *Homogeneous annealing*. We assign to each level of the pyramid the same, initially high, temperature. The relaxation is performed with this fixed temperature until an equilibrium is reached (i.e., until the change of the energy function associated with the model is less than a threshold). The temperature is then lowered. The algorithm is stopped at a low temperature close to 0. This algorithm can be described by a sequence of homogeneous Markov chains which are generated at a fixed temperature. The temperature will be decreased in between subsequent Markov chains.

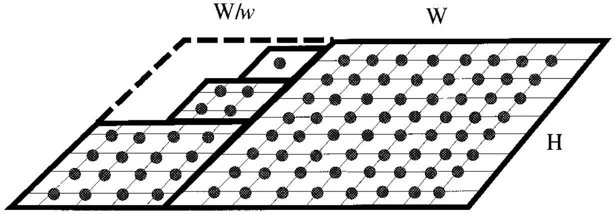


FIG. 7. Memory complexity of the hierarchical model.

2. *Inhomogeneous annealing*. The same initially high temperature is assigned to each level; however, the temperature is now lowered after each transition. In this case, the algorithm is described by an inhomogeneous Markov chain where the temperature is decreased in between subsequent transitions.

3. *Multitemperature annealing* (MTA). To the higher levels, we associate higher temperatures which enable the algorithm to be less sensitive to local minima. However, at a finer resolution, the relaxation is performed at a lower temperature (at the bottom level, it is close to 0).

In all cases, the final configuration of the finest level is taken as the solution of the problem.

4.2. Complexity

In this section, we study the complexity of the optimization of the hierarchical model in terms of the required memory (or number of processors in the parallel implementation) and the required communication compared to the monogrid model.

Memory/processor. We refer to the notations of the Section 2: let us suppose that our image is of the size $W \times H$. Following the procedure described in Section 2, we generate a pyramid containing $M + 1$ levels. Without loss of generality, we can assume that $W/w \leq H/h$, where $w \times h$ is the block size and both w and h are greater than or equal to two. The hierarchical model requires a maximum of $(1 + 1/w)WH$ processors (cf. Eq. (33)), since all levels must be stored at the same time. The memory (or processors) required for the storage of these levels (see Fig. 7), considering a rectangular shape, is given by

$$\begin{aligned} WH + \frac{WH}{wh} + \frac{WH}{(wh)^2} + \cdots + \frac{WH}{(wh)^M} \\ = WH \sum_{i=0}^{M} \frac{1}{(wh)^i} < \left(1 + \frac{1}{w}\right) WH. \end{aligned} \tag{33}$$

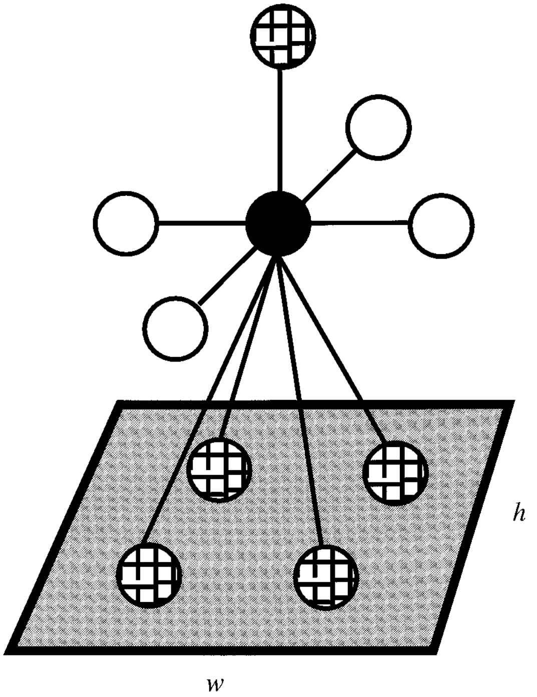


FIG. 8. Communication scheme of the hierarchical model.

Communication. Considering only the first- and second-order cliques (mostly used in practice, see Fig. 8), it is clear that we have (wh + 1) more communications per processors. Each site interacts with its ancestor (there is one) and its descendants (there are wh).

It turns out that the new model demands more processors and more computer time. However, as we can see later, experiments show that the new interaction is a better way to communicate between the grids yielding faster convergence (with respect to the number of iterations) for the stochastic relaxation algorithms and giving estimates which are closer to the global optimum for deterministic as well as for stochastic relaxation schemes.

4.2.1. *Multi-Temperature Annealing*

The main purpose and study of this section is a new MTA schedule. In this case, the configurations are generated at different temperatures at different sites. The temperature is then lowered after each transition according to the MTA schedule (see Theorem 4.1). More generally, we have the following problem:

Let $\mathcal{S} = \{s_1, \dots, s_N\}$ be a set of sites, $\mathcal{G}$ some neighborhood system with cliques $\mathcal{C}$, and X an MRF over these sites with energy function U. We define an annealing scheme where the temperature T depends on the iteration k and on the cliques C. Let $\oslash$ denote the operation

$$P(X = \omega) = \pi_{T(k,C)}(\omega) = \frac{\exp(-U(\omega) \oslash T(k, C))}{Z}, \quad (34)$$

where

$$U(\omega) \oslash T(k, C) = \sum_{C \in \mathcal{C}} \frac{V_C(\omega)}{T(k, C)}. \quad (35)$$

Let us suppose that the sites are visited for updating in the order $\{n_1, n_2, \dots\} \subset \mathcal{S}$. The resulting stochastic process is denoted by $\{X(k),\ k = 0, 1, 2, \dots\}$, where $X(0)$ is the initial configuration. $X(k)$ is an inhomogeneous Markov chain with transition matrix

$$P_{\omega,\eta}(k-1, k) = \begin{cases} G_{\omega,\eta}(T(k, C))A_{\omega,\eta}(T(k, C)) & \forall \eta \neq \omega \\ 1 - \sum_{\zeta \neq \omega} G_{\omega,\zeta}(T(k, C))A_{\omega,\zeta}(T(k, C)) & \eta = \omega. \end{cases} \quad (36)$$

Considering the Gibbs sampler, the generation matrix $G_{\omega,\eta}(T(k, C))$ and acceptation matrix $A_{\omega,\eta}(T(k, C))$ is given by

$$G_{\omega,\eta}(T(k, C)) = G_{\omega,\eta}(k) = \begin{cases} 1, & \text{if } \eta = \omega|_{\omega_{n_k}=\lambda} \text{ for some } \lambda \in \Lambda \\ 0, & \text{otherwise} \end{cases} \quad (37)$$

$$A_{\omega,\eta}(T(k, C)) = \pi_{T(k,C)}(X_{n_k} = \omega_{n_k} \mid X_s = \omega_s, s \neq n_k). \quad (38)$$

Note that the acceptance is governed by the *local* characteristics. $\pi_{T(k,C)}(X_{n_k} = \omega_{n_k} \mid X_s = \omega_s,\ s \neq n_k)$ has a slightly different meaning than $\pi_{T(k,C)}(\omega)$ in Eq. (34):

$$\pi_{T(k,C)}(X_s = \omega_s \mid X_r = \omega_r, s \neq r) = \frac{1}{Z_s} \exp\left(-\frac{\sum_{C \in \mathcal{C}: s \in C} V_C(\omega)}{T(k, C)}\right) \quad (39)$$

with

$$Z_s = \sum_{\lambda \in \Lambda} \exp\left(-\frac{\sum_{C \in \mathcal{C}: s \in C} V_C(\omega|_{\omega_s=\lambda})}{T(k, C)}\right). \quad (40)$$

The transition matrix at time k is then of the form

$$P_{\omega,\eta}(k) = \begin{cases} \pi_{T(k,C)}(X_{n_k} = \eta_{n_k} \mid X_s = \eta_s, s \neq n_k), \\ \quad \text{if } \eta = \omega|_{\omega_{n_k}=\lambda} \text{ for some } \lambda \in \Lambda \\ 0, \\ \quad \text{otherwise.} \end{cases} \quad (41)$$

Let Ω_{opt} be the set of globally optimal configurations

$$\Omega_{\text{opt}} = \left\{\omega \in \Omega : U(\omega) = \min_{\eta \in \Omega} U(\eta)\right\}. \tag{42}$$

Let π_0 be the uniform distribution on Ω_{opt}, and define

$$U^{\sup} = \max_{\omega \in \Omega} U(\omega), \tag{43}$$

$$U^{\inf} = \min_{\omega \in \Omega} U(\omega), \tag{44}$$

and

$$\Delta = U^{\sup} - U^{\inf}. \tag{45}$$

Let us examine the decomposition of $U(\omega) \oslash T(k, C)$ defined in Eq. (35). Let $\omega' \in \Omega_{\text{opt}}$ be a *globally* optimal configuration. Thus, $U(\omega') - U^{\inf}$ equals 0. In the case of a classical annealing, dividing by a constant temperature does not change this relation (obviously, $\forall k$: $(U(\omega') - U^{\inf})/T_k$ is still 0). But it is not necessarily true that $(U(\omega') - U^{\inf}) \oslash T(k, C)$ is also 0! Because choosing sufficiently small temperatures for the cliques where ω'_C is locally not optimal (i.e., strengthening the nonoptimal cliques) and choosing sufficiently high temperatures for the cliques where ω'_C is locally optimal (i.e., weakening the optimal cliques), we obtain $(U(\omega') - U^{\inf}) \oslash T(k, C) > 0$, meaning that ω' is no longer globally optimal.

Thus, we must impose further conditions on the temperature to assure the convergence. First, let us examine the decomposition over the cliques of $U(\omega) - U(\eta)$ for arbitrary ω and η, $\omega \neq \eta$:

$$U(\omega) - U(\eta) = \sum_{C \in \mathscr{C}} (V_C(\omega) - V_C(\eta)). \tag{46}$$

Indeed, there may be negative and positive members in the decomposition. According to this fact, we have the subsums

$$\begin{aligned} \sum_{C \in \mathscr{C}} (V_C(\omega) - V_C(\eta)) &= \underbrace{\sum_{C \in \mathscr{C} : (V_C(\omega) - V_C(\eta)) < 0} (V_C(\omega) - V_C(\eta))}_{\Sigma^{-}(\omega, \eta)} \\ &\quad + \underbrace{\sum_{C \in \mathscr{C} : (V_C(\omega) - V_C(\eta)) \geq 0} (V_C(\omega) - V_C(\eta)).}_{\Sigma^{+}(\omega, \eta)} \end{aligned} \tag{47}$$

Now, let us examine Δ defined in Eq. (45). If we want to decompose Δ as defined above, we must choose some configuration ω' with a maximum energy (i.e., $U(\omega') = U^{\sup}$) and another configuration ω'' with a minimum energy (i.e., $U(\omega'') = U^{\inf}$). Obviously, there may be more than one decomposition, depending on the number of globally optimal configurations ($|\Omega_{\text{opt}}|$) and the number of configurations with maximal global energy ($|\Omega_{\sup}|$). Thus, the decomposition of Δ for a given (ω', ω'') is of the form

$$\Delta = \Sigma^{-}(\omega', \omega'') + \Sigma^{+}(\omega', \omega'') \tag{48}$$

Furthermore, let us define Σ^{+}_{Δ} as

$$\Sigma^{+}_{\Delta} = \min_{\substack{\omega' \in \Omega_{\sup} \\ \omega'' \in \Omega_{\text{opt}}}} \Sigma^{+}(\omega', \omega''). \tag{49}$$

Obviously, $\Delta \leq \Sigma^{+}_{\Delta}$. The following theorem gives an annealing schedule, basically the same as in [11]. *However, the temperature here is a function of k and $C \in \mathscr{C}$.*

THEOREM 4.1 (MULTITEMPERATURE ANNEALING). *Assume that there exists an integer $\kappa \geq N$ such that for every $k = 0, 1, 2, \ldots, \mathscr{S} \subseteq \{n_{k+1}, n_{k+2}, \ldots, n_{k+\kappa}\}$. For all $C \in \mathscr{C}$, let $T(k, C)$ be any decreasing sequence of temperatures in k for which*

1. $\lim_{k \to \infty} T(k, C) = 0$. *Let us denote respectively by $T^{\inf}_k$ and $T^{\sup}_k$ the maximum and minimum of the temperature function at k ($\forall C \in \mathscr{C}$: $T^{\inf}_k \leq T(k, C) \leq T^{\sup}_k$).*

2. *For all $k \geq k_0$, for some integer $k_0 \geq 2$: $T^{\inf}_k \geq N\Sigma^{+}_{\Delta}/\ln(k)$.*

3. *If $\Sigma^{-}(\omega, \omega') \neq 0$ for some $\omega \in \Omega \backslash \Omega_{\text{opt}}$, $\omega' \in \Omega_{\text{opt}}$ then a further condition must be imposed*:

For all k: $(T^{\sup}_k - T^{\inf}_k)/T^{\inf}_k \leq R$ with

$$R = \min_{\substack{\omega \in \Omega/\Omega_{\text{opt}} \\ \omega' \in \Omega_{\text{opt}} \\ \Sigma^{-}(\omega, \omega' \neq 0}} \frac{U(\omega) - U^{\inf}}{|\Sigma^{-}(\omega, \omega')|}. \tag{50}$$

Then for any starting configuration $\eta \in \Omega$ and for every $\omega \in \Omega$,

$$\lim_{k \to \infty} P(X(k) = \omega \mid X(0) = \eta) = \pi_0(\omega). \tag{51}$$

The proof of this theorem appears in the Appendix.

Remarks.

1. In practice, we cannot determine R and Σ^{+}_{Δ}, nor can we compute Δ.

2. Considering Σ^{+}_{Δ} in condition 2, we have the same problem as in the case of a classical annealing. The only difference is that in a classical annealing, we have Δ instead of Σ^{+}_{Δ}. Consequently, the same solutions may be used: an exponential schedule with a sufficiently high initial temperature.

TABLE 1
Results on a Noisy Synthetic Image with Four Classes

	Levels	VPR	T_0	Iterations	Total time (s)	Time/iteration (s)	Error	β	γ
					Original				
Gibbs	1	2	4	68	3.01	0.04	183 (1.12%)	1.0	—
ICM	1	2	1	9	0.15	0.02	2948 (17.99%)	1.0	—
					Multiscale				
Gibbs	4	1,2	4	101	3.85	0.04	176 (1.07%)	1.0	—
ICM	4	1,2	1	17	0.22	0.01	1657 (10.11%)	0.9	—
					Hierarchical				
Gibbs	4	4	4,3,2,1	41	141.97	3.46	191 (1.16%)	0.7	0.1
ICM	4	4	1	11	30.17	2.74	293 (1.78%)	0.8	0.5

3. The factor R is more interesting. We propose herein two possibilities which can be used for practical implementations of the method: Either we choose a sufficiently small interval $[T_0^{\text{inf}}, T_0^{\text{sup}}]$ and suppose that it satisfies condition 3 (we have used this technique in the simulations) or we use a more strict but easily verifiable condition instead of condition 3, namely,

$$\lim_{k\to\infty} \frac{T_k^{\text{sup}} - T_k^{\text{inf}}}{T_k^{\text{inf}}} = 0. \tag{52}$$

4. What happens if $\Sigma^-(\omega, \omega')$ is zero for all ω and ω' in condition 3 and thus R is not defined? This is the best case because it means that all globally optimal configurations are also locally optimal. That is, we have no restriction on the interval $[T_k^{\text{inf}}, T_k^{\text{sup}}]$; thus, any local temperature schedule satisfying conditions 1 and 2 is good.

5. EXPERIMENTAL RESULTS

We compare the Gibbs sampler [11] and iterated conditional mode [3, 18] using three models for each algorithm (original, multiscale, and hierarchical). We have also compared the inhomogeneous and MTA schedules. All tests have been conducted on a Connection Machine CM200 [16] with $8K$ processors. In Tables 1 and 2, we give for each model and for each method the number of levels in the pyramid (for the monogrid model, this is 1), the virtual processor ratio (VPR) [16], the initial temperature (for the hierarchical model, this is not the same at each level, using the MTA schedule), the number of iterations, the computing time, the error of the classification (= the number of misclassified pixels), and the parameter β (see Eqs. (3), (15), (16)) and γ (see Eq. (32)).

5.1. Comparison of the Schedules

In Fig. 11 we compare the inhomogeneous and MTA schedules on a noisy synthetic image using the Gibbs sampler. In both cases, the parameters were strictly the same, the only difference is the applied schedule: the pyramid contains four levels yielding a VPR equal to 4. The initial temperature were respectively 4 (at the highest level), 3, 2, and 1 (at the lowest level). The potential β equals 0.7 and γ equals 0.1. In Fig. 10 (resp. Fig. 9), we show the global energy (computed at a fixed temperature) versus the number of iterations of the inhomogeneous (resp. MTA) schedule. Both reach practically the same minimum (53415.4 for the inhomogeneous and 53421.4 for the MTA); however, the inhomogeneous schedule requires 238 iterations (796.8 s CPU time) but the MTA schedule requires only 100 iterations (340.6 s CPU time) for the convergence.

5.2. Comparison of the Models

First, we tested the models on a noisy synthetic image of size 128 × 128. In the image, we have different geometrical forms (circle and triangle) on a checkerboard image (see Fig. 12, Table 1). The Gibbs sampler gives nearly the same result in all cases. However the ICM is more sensitive. The multiscale model gives better result than the monogrid one but the result is not fine in the triangle and the circle. These forms have a different structure than the block structure of the model, the initialization was wrong in these regions, and the ICM was not able to correct these errors. In the hierarchical case, instead of the initialization, we have a real time communication between the levels which is able to give results close to those obtained with the Gibbs sampler. This model requires quite greater computing time than the others. The reason is that, in the hierarchical case, the whole pyramid is stored at the same time, yielding a greater VPR ratio. On the other hand, we cannot use the fast “NEWS” communication scheme [16] as in the other cases.

Finally, we present a SPOT image of size 512 × 512 (see Fig. 13) with ground truth data. In the following table, we

TABLE 2
Results on the SPOT Image with Six Classes

	Levels	VPR	T_0	Iterations	Total time (s)	Time/iteration (s)	β	γ
				Original				
Gibbs	1	32	4	234	163.18	0.69	1.5	—
ICM	1	32	1	8	2.03	0.25	1.5	—
				Multiscale				
Gibbs	5	1–32	4	580	180.17	0.31	1.5	—
ICM	5	1–32	1	36	5.15	0.14	0.3	—
				Hierarchical				
Gibbs	5	64	4,3,2,1	154	9629.33	62.53	0.7	0.1
ICM	5	64	1	16	915.99	57.25	1.0	0.2

give the mean (μ) and the deviation (σ^2) for each class (we have six classes):

Class	1	2	3	4	5	6
μ	65.3	81.3	75.4	98.5	82.5	129.0
σ^2	6.4	12.7	14.9	16.8	9.46	183.2

As we can see, classes 2 and 5 have nearly the same parameters; it is difficult to distinguish between them. Figure 14 (resp. Fig. 15) shows the results obtained with the ICM (resp. Gibbs sampler). For these results, we give a map drawn by an expert (ground truth data). Classes 1–6 correspond to the regions B_{3c}, B_{3b}, B_{3d}, a_2, hc, and 92_a on the map. For the hierarchical model a slight improvement can be noticed for the results of the Gibbs sampler; however, for the ICM, the improvement is more significant. In Table 2 we give the parameters and the computing time for each model and each method.

6. CONCLUSION

In this paper, we have presented a classical multiscale model and proposed a new hierarchical MRF model. We have introduced a new interaction scheme between two neighbor grids in the label pyramid and have experimen-

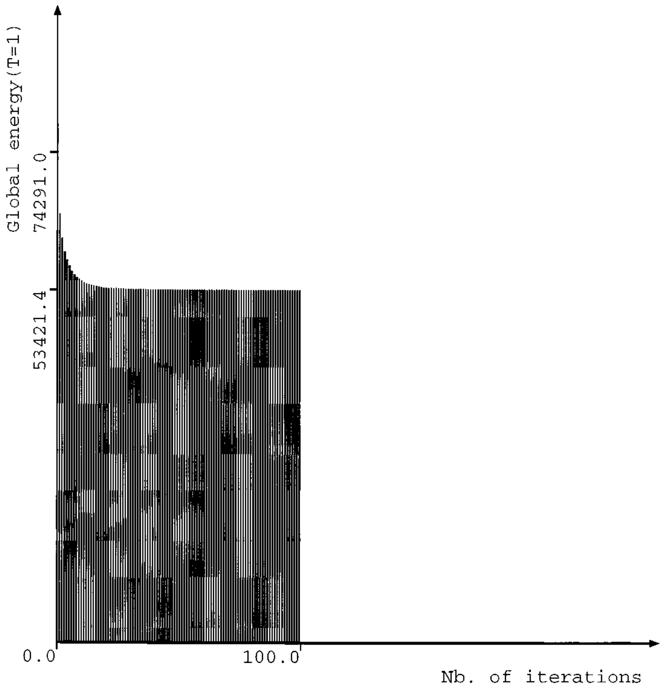


FIG. 9. Energy decrease with the MTA schedule.

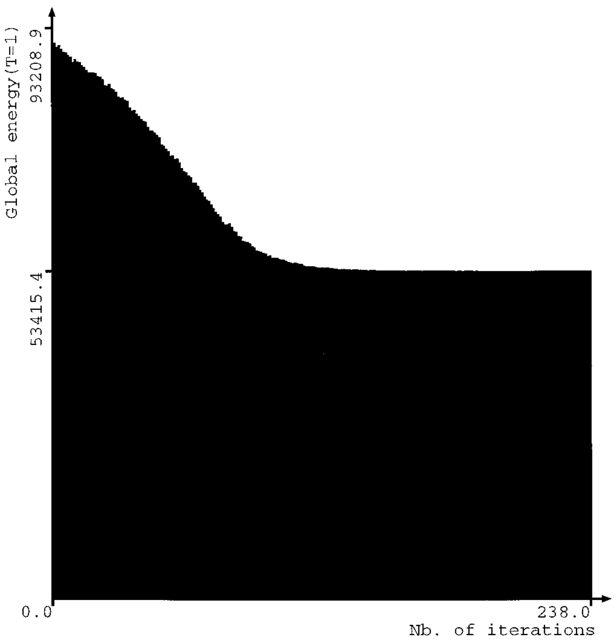
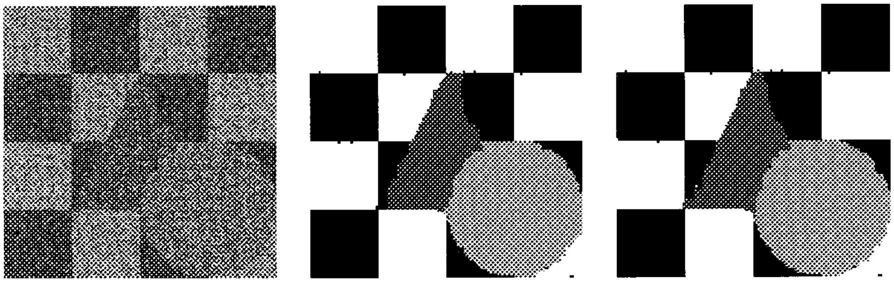


FIG. 10. Energy decrease with the inhomogeneous annealing schedule.

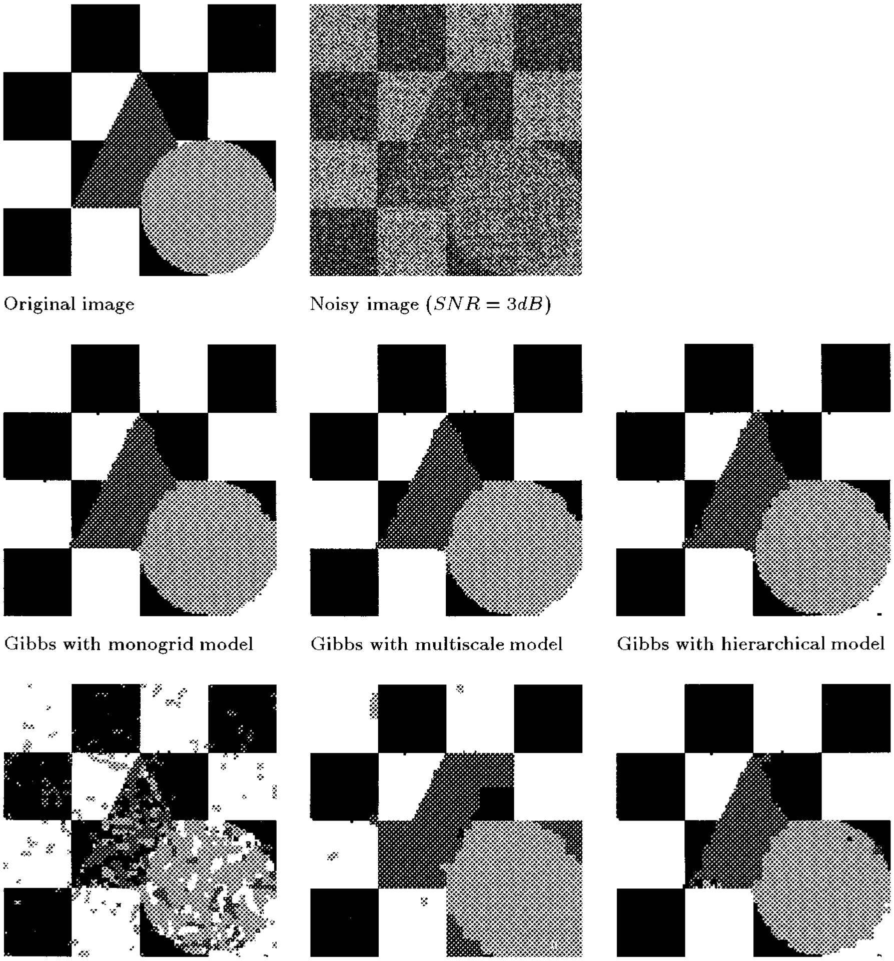


Noisy image ($SNR = 3dB$) Inhomogeneous schedule MTA schedule

FIG. 11. Results of the Gibbs sampler on a synthetic image with inhomogeneous and MTA schedules.

tally shown that these connections allow us to propagate local interactions more efficiently, yielding faster convergence (w.r.t. the number of iterations) in many cases and giving estimates closer to the optimum for deterministic as well as for stochastic relaxation techniques. On the other hand, these interactions make the model more complex, demanding computationally more expensive algorithms.

We have also proposed a new general annealing scheme, the multitemperature annealing. We have used



Original image Noisy image ($SNR = 3dB$)

Gibbs with monogrid model Gibbs with multiscale model Gibbs with hierarchical model

ICM with monogrid model ICM with multiscale model ICM with hierarchical model

FIG. 12. Results on a synthetic image with four classes.

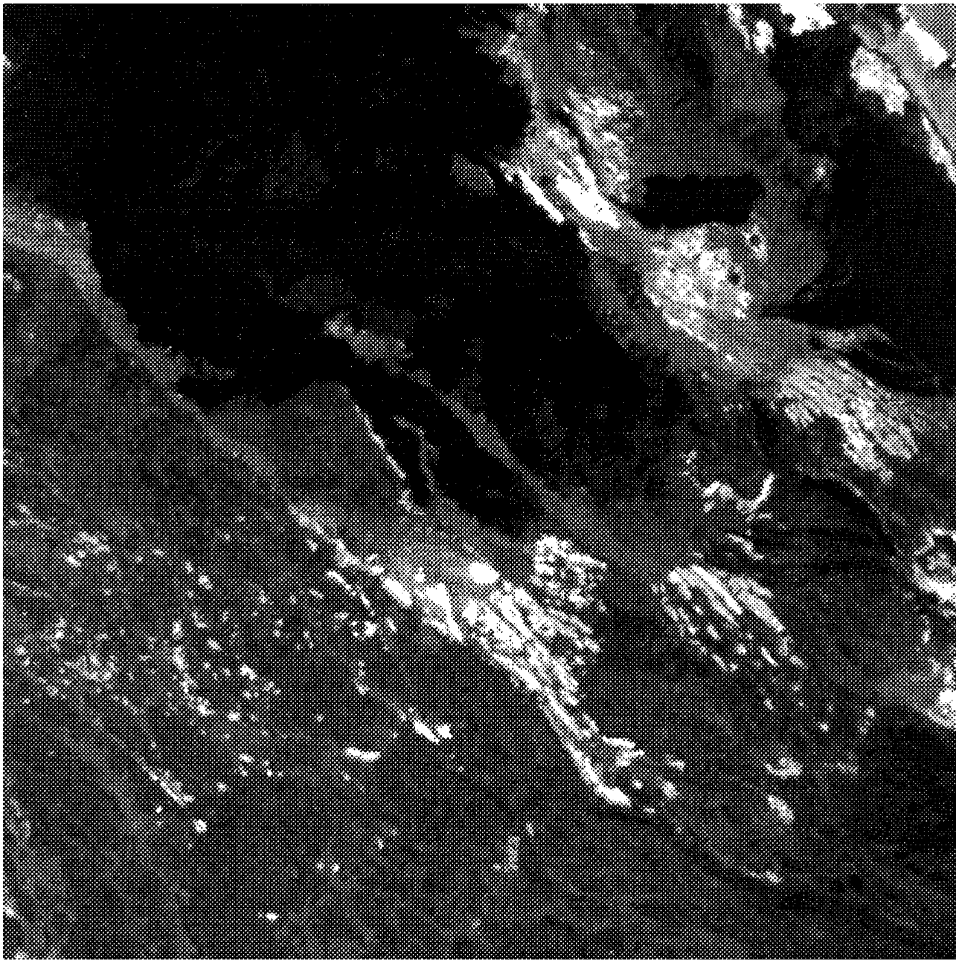


FIG. 13. Original SPOT image with six classes.

a degenerated case of this MTA scheme for the minimization of the energy function of the hierarchical model: the temperature decreasing scheme is rigid with different fixed coefficients applied to the different levels of the label pyramid. This algorithm can be run in parallel on the entire pyramid and usually decreases the computational time compared to the classical schemes. A generalization of the annealing theorem of Geman and Geman [11] has been proposed, which gives a theoretical background for the convergence of this method toward global optimum.

Finally, the hierarchical model and the theoretical study given in this paper are presented in a general form. Although they have been adapted for supervised image classification, one can also use them for other low level vision tasks such as edge detection, image restoration, data fusion, motion, etc. We are currently working on the parameter estimation of these models for unsupervised image classification.

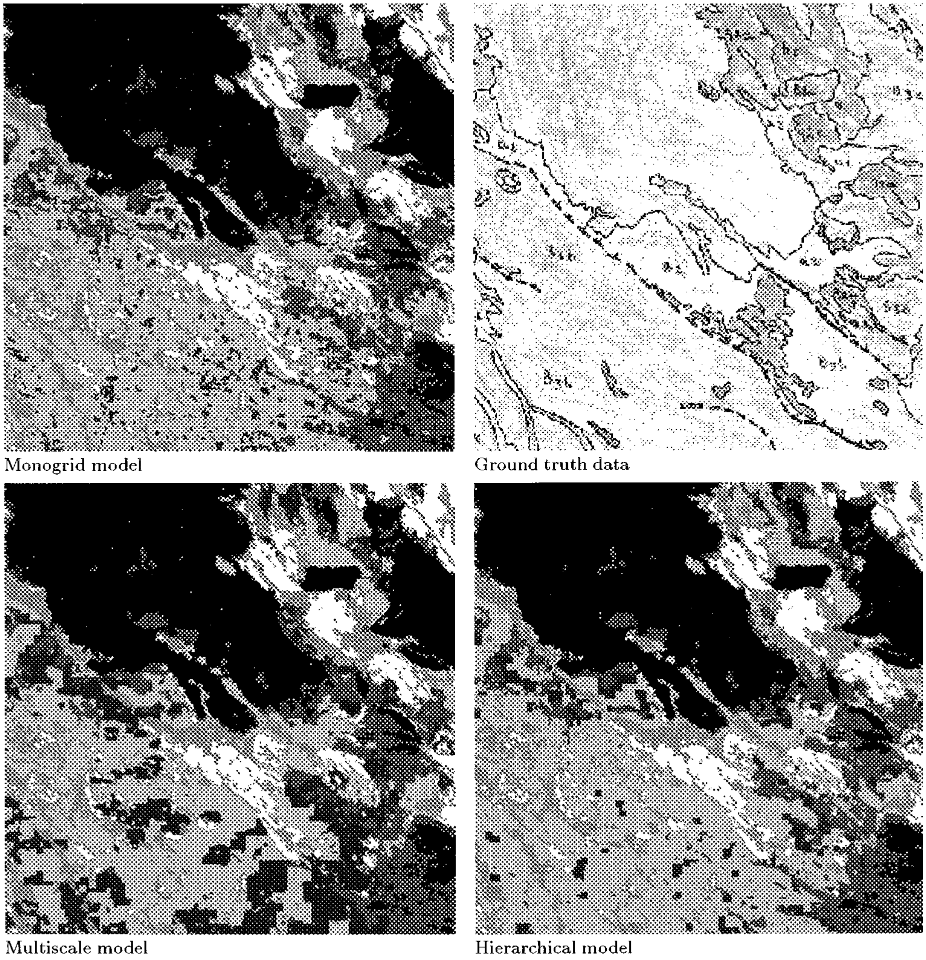
APPENDIX

Proof of the Multitemperature Annealing Theorem

We follow the proof of the annealing theorem given by Geman and Geman in [11]. Essentially, we can apply the same proof, only a slight modification is needed.

A.1. *Notations*

We recall a few notations: $\mathcal{S} = \{s_1, \ldots, s_N\}$ denotes the set of sites, $\Lambda = \{0, 1, \ldots, L - 1\}$ is a common state space, and $\omega, \eta, \eta' \ldots \in \Omega$ denote configurations, where $\Omega = \Lambda^N$ is finite. The sites are updated in the order $\{n_1, n_2, \ldots\} \subset \mathcal{S}$. The generated configurations constitute an inhomogeneous Markov chain $\{X(k), k = 0, 1, 2, \ldots\}$, where $X(0)$ is the initial configuration. The transition $X(k - 1) \to X(k)$ is controlled by the Gibbs distribution $\pi_{T(k,C)}$ according to the transition matrix at time k:



Monogrid model

Ground truth data

Multiscale model

Hierarchical model

FIG. 14. Results of the ICM algorithm. Comparison with ground truth data.

$$P_{\omega,\eta}(k) = \begin{cases} \pi_{T(k,C)}(X_{n_k} = \eta_{n_k} \,|\, X_s = \eta_s, s \neq n_k), \\ \quad \text{if } \eta = \omega|_{\omega_{n_k}=\lambda} \text{ for some } \lambda \in \Lambda \\ 0, \\ \quad \text{otherwise.} \end{cases} \tag{53}$$

$\pi_{T(k,C)}(\omega)$ denotes the Gibbs distribution at iteration k

$$\pi_{T(k,C)}(\omega) = \frac{\exp(-U(\omega) \oslash T(k,C))}{Z} \tag{54}$$

with

$$U(\omega) \oslash T(k,C) = \sum_{C \in \mathscr{C}} \frac{V_C(\omega)}{T(k,C)}. \tag{55}$$

The local characteristics of the above distribution are denoted by

$$\begin{aligned} &\pi_{T(k,C)}(X_s = \omega_s \,|\, X_r = \omega_r, s \neq r) \\ &\quad = \frac{1}{Z_s} \exp\left(-\frac{\sum_{C \in \mathscr{C}: s \in C} V_C(\omega)}{T(k,C)} \right) \end{aligned} \tag{56}$$

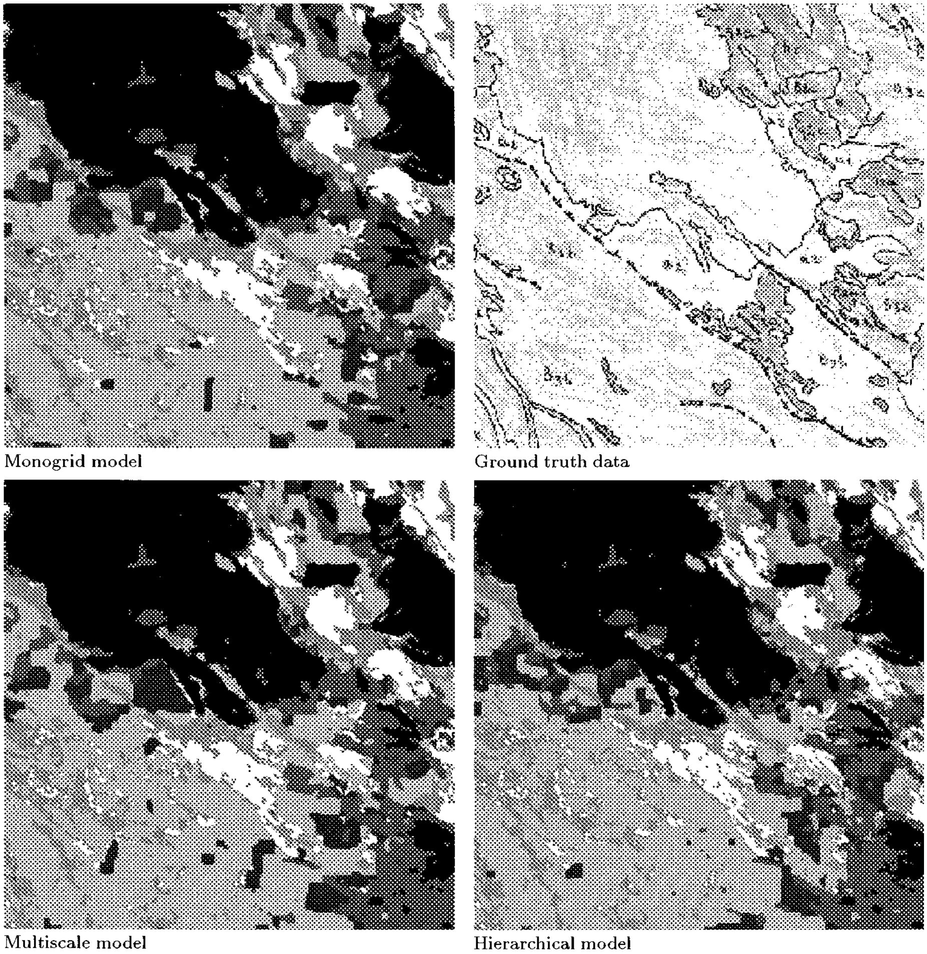
with

$$Z_s = \sum_{\lambda \in \Lambda} \exp\left(-\frac{\sum_{C \in \mathscr{C}: s \in C} V_C(\omega|_{\omega_s=\lambda})}{T(k,C)} \right). \tag{57}$$

The decomposition of $U(\omega) - U(\eta)$ for arbitrary ω and η, $\omega \neq \eta$ is given by

$$U(\omega) - U(\eta) = \sum_{C \in \mathscr{C}} (V_C(\omega) - V_C(\eta)). \tag{58}$$

Denoting respectively by $\sum^{+}(\omega, \eta)$ and $\sum^{-}(\omega, \eta)$ the sum over the positive and negative cliques, we get



Monogrid model

Ground truth data

Multiscale model

Hierarchical model

FIG. 15. Results of the Gibbs sampler. Comparison with ground truth data.

$$\begin{aligned}\sum_{C\in\mathscr{C}} (V_C(\omega) - V_C(\eta)) &= \underbrace{\sum_{C\in\mathscr{C}:(V_C(\omega)-V_C(\eta))<0} (V_C(\omega) - V_C(\eta))}_{\Sigma^-(\omega,\eta)} \\ &\quad + \underbrace{\sum_{C\in\mathscr{C}:(V_C(\omega)-V_C(\eta))\geq 0} (V_C(\omega) - V_C(\eta)).}_{\Sigma^+(\omega,\eta)}\end{aligned} \tag{59}$$

Furthermore, let

$$U^{\sup} = \max_{\omega\in\Omega} U(\omega), \tag{60}$$

$$U^{\inf} = \min_{\omega\in\Omega} U(\omega), \tag{61}$$

and

$$\Delta = U^{\sup} - U^{\inf}. \tag{62}$$

and define Σ_Δ^+ as the minimum of positive sums:

$$\Sigma_\Delta^+ = \min_{\substack{\omega'\in\Omega_{\text{sup}} \\ \omega''\in\Omega_{\text{opt}}}} \Sigma^+(\omega', \omega''). \tag{63}$$

Obviously, $\Delta \leq \Sigma_\Delta^+$.

Given any starting distribution μ_0, the distribution of $X(k)$ is given by the vector $\mu_0 \prod_{i=1}^k P(i)$:

$$P_{\mu_0}(X(k) = \omega) = \left(\mu_0 \prod_{i=1}^{k} P(i)\right)\Bigg|_{\omega} \tag{64}$$

$$= \sum_{\eta} P(X(k) = \omega \mid X(0) = \eta)\mu_0(\eta). \tag{65}$$

We use the following notation for transitions: $\forall l < k$ and ω, $\eta \in \Omega$,

$$P(k, \omega \,|\, l, \eta) = P(X(k) = \omega \,|\, X(l) = \eta),$$

and for any distribution μ on Ω,

$$P(k, \omega \,|\, l, \mu) = \sum_{\eta} P(X(k) = \omega \,|\, X(l) = \eta)\mu(\eta).$$

Sometimes, we use this notation as $P(k, \cdot \,|\, l, \mu)$, where "$\cdot$" means *any* configuration from Ω. Finally, let $\|\mu - \nu\|$ denotes the following distance between two distributions on Ω:

$$\|\mu - \nu\| = \sum_{\omega} |\mu(\omega) - \nu(\omega)|.$$

It is clear that $\lim_{n\to\infty} \mu_n = \mu$ in distribution (i.e., $\forall\omega : \mu_n(\omega) \to \mu(\omega)$) if and only if $\|\mu_n - \mu\| \to 0$.

A.2. *Proof of the Theorem*

First, we state two lemmas which imply Theorem 4.1:

LEMMA A.1. *For every* $k_0 = 0, 1, 2 \ldots,$

$$\lim_{k\to\infty} \sup_{\omega,\eta',\eta''} |P(X(k) = \omega \,|\, X(k_0) = \eta') - P(X(k) = \omega \,|\, X(k_0) = \eta'')| = 0. \tag{66}$$

Proof of Lemma A.1. Fix $k_0 = 0, 1, 2, \ldots,$ define $K_l = k_0 + l\kappa, l = 0, 1, 2, \ldots,$ where κ is the number of transitions necessary for a full sweep of $\mathscr{S}$ (for every $k = 0, 1, 2, \ldots, \mathscr{S} \subseteq \{n_{k+1}, n_{k+2}, \ldots, n_{k+\kappa}\}$). Let $\delta(k)$ be the smallest probability among the local characteristics:

$$\delta(k) = \inf_{\substack{1\le i\le N\\ \omega\in\Omega}} \pi_{T(k,C)}(X_{s_i} = \omega_{s_i} \,|\, X_{s_j} = \omega_{s_j}, j \neq i).$$

A lower bound for $\delta(k)$ is given by

$$\delta(k) \geq \frac{\exp(-U^{\sup} \oslash T(k, C))}{L \exp(-U^{\inf} \oslash T(k, C))} = \frac{\exp(-\Delta \oslash T(k, C))}{L}$$
$$\geq \frac{1}{L} \exp(-\Sigma_{\Delta}^{+} \oslash T(k, C) \geq \frac{1}{L} \exp(-\Sigma_{\Delta}^{+}/T_k^{\inf}),$$

where $L = |\Lambda|$ is the number of possible states at a site. Now fix l and define m_i as the time of the last replacement of site s_i before $K_l + 1$ (that is, before the lth full sweep):

$$\forall i \colon 1 \leq i \leq N \colon m_i = \sup\{k : k \leq K_l, n_k = s_i\}.$$

Without loss of generality, we can assume that $m_1 > m_2 \cdots > m_N$ (otherwise, relabel the sites). Then

$$\begin{aligned} P(X(K_l) = \omega \,|\, X(K_{l-1}) = \omega') &= P(X_{s_1}(m_1) = \omega_{s_1}, X_{s_2}(m_2) = \omega_{s_2}, \ldots, X_{s_N}(m_N) = \omega_{s_N} \,|\, X(K_{l-1}) = \omega') \\ &= \prod_{i=1}^{N-1} P(X_{s_i}(m_i) = \omega_{s_i} \,|\, X_{s_{i+1}}(m_{i+1}) = \omega_{s_{i+1}}, \ldots, X_{s_N}(m_N) = \omega_{s_N}, X(K_{l-1}) = \omega') \qquad (67) \\ &\geq \prod_{i=1}^{N} \delta(m_i) \geq L^{-N} \prod_{i=1}^{N} \exp(-\Delta/T_{m_i}^{\inf}) \geq L^{-N} \exp\left(-\frac{\Sigma_{\Delta}^{+} N}{T_{k_0+l_\kappa}^{\inf}}\right), \end{aligned}$$

since $m_i \leq K_l = k_0 + l\kappa$, $i = 1, 2 \ldots, N$ and $T_k^{\inf}$ is decreasing. If $k_0 + l\kappa$ is sufficiently large then $T_{k_0+l\kappa}^{\inf} \geq N \, \Sigma_{\Delta}^{+}/\ln(k_0 + l\kappa)$ according to condition 2, and Eq. (67) can be continued as

$$P(X(K_l) = \omega \,|\, X(K_{l-1}) = \omega')$$
$$\geq L^{-N} \exp\left(-\frac{\Sigma_{\Delta}^{+} N}{N \, \Sigma_{\Delta}^{+}/\ln(k_0 + l\kappa)}\right) = L^{-N}(k_0 + l\kappa)^{-1}.$$

Hence, for a sufficiently small constant $\Gamma (0 < \Gamma \leq 1)$, we can assume that

$$\inf_{\omega,\omega'} P(X(K_l) = \omega \,|\, X(K_{l-1}) = \omega') \geq \frac{\Gamma L^{-N}}{k_0 + l\kappa} \tag{68}$$

for every $k_0 = 0, 1, 2, \ldots$ and $l = 1, 2, \ldots,$ keeping in mind that K_l depends on k_0.

Consider now the limit given in Eq. (66) and for each $k > k_0$, define $K^{\sup}(k) = \sup\{l : K_l < k\}$ (the last sweep before the kth transition) so that $\lim_{k\to\infty} K^{\sup}(k) = \infty$. Fix $k > K_1$:

$$\sup_{\omega,\eta',\eta''} |P(X(k) = \omega \,|\, X(0) = \eta') - P(X(k) = \omega \,|\, X(0) = \eta'')|$$
$$= \sup_{\omega} \big(\sup_{\eta} P(X(k) = \omega \,|\, X(0) = \eta)$$
$$- \inf_{\eta} P(X(k) = \omega \,|\, X(0) = \eta)\big)$$
$$= \sup_{\omega} \Bigg(\sup_{\eta} \sum_{\omega'} P(X(k) = \omega \,|\, X(K_1) = \omega')$$
$$P(X(K_1) = \omega' \,|\, X(0) = \eta)$$
$$- \inf_{\eta} \sum_{\omega'} P(X(k) = \omega \,|\, X(K_1) = \omega')$$

$$P(X(K_1) = \omega' \,|\, X(0) = \eta)\Big)$$

$$\doteq \sup_{\omega} Q(k, \omega).$$

Furthermore, for each $\omega \in \Omega$,

$$\sup_{\eta} \sum_{\omega'} P(X(k) = \omega \,|\, X(K_1) = \omega')$$
$$P(X(K_1) = \omega' \,|\, X(0) = \eta)$$
$$\leq \sup_{\mu} \sum_{\omega'} P(X(k) = \omega \,|\, X(K_1) = \omega')\mu(\omega'),$$

where μ is any probability measure on Ω. Using Eq. (68), we get

$$\mu(\omega') \geq \frac{\Gamma L^{-N}}{k_0 + l\kappa}.$$

Suppose that $P(X(k) = \omega \,|\, X(K_1) = \omega')$ is maximized at $\omega' = \omega^{\sup}$ and minimized at $\omega' = \omega^{\inf}$. Then we get

$$\sup_{\mu} \sum_{\omega'} P(X(k) = \omega \,|\, X(K_1) = \omega')\mu(\omega')$$
$$\leq \left(1 - (L^N - 1)\frac{\Gamma L^{-N}}{k_0 + l\kappa}\right) P(X(k) = \omega \,|\, X(K_1) = \omega^{\sup})$$
$$+ \frac{\Gamma L^{-N}}{k_0 + l\kappa} \underbrace{\sum_{\omega' \neq \omega^{\sup}} P(X(k) = \omega \,|\, X(K_1) = \omega'),}_{P(X(k)=\omega \,|\, X(K_1)=\omega^{\inf}) + \sum_{\omega' \neq \omega^{\sup},\omega^{\inf}} P(X(k)=\omega|_{X}(K_1)=\omega'),}$$

and in a similar way

$$\inf_{\mu} \sum_{\omega'} P(X(k) = \omega \,|\, X(K_1) = \omega')\mu(\omega')$$
$$\geq \left(1 - (L^N - 1)\frac{\Gamma L^{-N}}{k_0 + l\kappa}\right) P(X(k) = \omega \,|\, X(K_1) = \omega^{\inf})$$
$$+ \frac{\Gamma L^{-N}}{k_0 + l\kappa} \underbrace{\sum_{\omega' \neq \omega^{\inf}} P(X(k) = \omega \,|\, X(K_1) = \omega').}_{P(X(k)=\omega \,|\, X(K_1)=\omega^{\sup}) + \sum_{\omega' \neq \omega^{\sup},\omega^{\inf}} P(X(k)=\omega \,|\, X(K_1)=\omega').}$$

Then it is clear that

$$Q(k, \omega) \leq \left(1 - \frac{\Gamma}{k_0 + l\kappa}\right)(P(X(k) = \omega \,|\, X(K_1) = \omega^{\sup})$$
$$- P(X(k) = \omega \,|\, X(K_1) = \omega^{\inf}));$$

hence,

$$\sup_{\omega,\eta',\eta''} |P(X(k) = \omega \,|\, X(0) = \eta') - P(X(k) = \omega \,|\, X(0) = \eta'')|$$
$$\leq \left(1 - \frac{\Gamma}{k_0 + l\kappa}\right) \sup_{\omega,\eta',\eta''} |P(X(k) = \omega \,|\, X(K_1) = \eta')$$
$$- P(X(k) = \omega \,|\, X(K_1) = \eta'')|$$
$$\leq \left(1 - \frac{\Gamma}{k_0 + l\kappa}\right)\left(\left(1 - \frac{\Gamma}{k_0 + l\kappa}\right) \sup_{\omega,\eta',\eta''} |P(X(k)\right.$$
$$\left. = \omega \,|\, X(K_2) = \eta') - P(X(k) = \omega \,|\, X(K_2) = \eta'')|\right).$$

Proceeding this way, we have the bound

$$\leq \prod_{k=1}^{K^{\sup}(k)} \left(1 - \frac{\Gamma}{k_0 + l\kappa}\right) \sup_{\omega,\eta',\eta''} |P(X(k) = \omega \,|\, X(K_{K^{\sup}(k)}) = \eta')$$
$$- P(X(k) = \omega \,|\, X(K_{K^{\sup}(k)}) = \eta'')|,$$

and finally, since the possible maximal value of the supremum is 1,

$$\sup_{\omega,\eta',\eta''} |P(X(k) = \omega \,|\, X(0) = \eta') - P(X(k) = \omega \,|\, X(0) = \eta'')|$$
$$\leq \prod_{k=1}^{K^{\sup}(k)} \left(1 - \frac{\Gamma}{k_0 + l\kappa}\right).$$

It is then sufficient to show that

$$\lim_{m \to \infty} \prod_{k=1}^{m} \left(1 - \frac{\Gamma}{k_0 + l\kappa}\right) = 0,$$

which is a well-known consequence of the divergence of the series

$$\sum_{l} (k_0 + l\kappa)^{-1}$$

for all k_0 and κ. This completes the proof of Lemma A.1. Q.E.D.

Lemma A.2.

$$\lim_{k_0 \to \infty} \sup_{k \geq k_0} \|P(k, \cdot \,|\, k_0, \pi_0) - \pi_0\| = 0. \tag{69}$$

Proof of Lemma A.2. In the following, let $P_{k_0,k}(\cdot)$ stand for $P(k, \cdot \,|\, k_0, \pi_0)$, so that for any $k \geq k_0 > 0$:

$$P_{k_0,k}(\omega) = \sum_{\eta} P(X(k) = \omega \,|\, X(k_0) = \eta)\pi_0(\eta).$$

First, we show that for any $k > k_0 \geq 0$:

$$\|P_{k_0,k} - \pi_{T(k,C)}\| \le \|P_{k_0,k-1} - \pi_{T(k,C)}\|. \tag{70}$$

We can assume for convenience that $n_k = s_1$. Then

$$\begin{aligned}
&\|P_{k_0,k} - \pi_{T(k,C)} \\
&\quad = \sum_{(\omega_{s_1},\dots,\omega_{s_N})} |\pi_{T(k,C)}(X_{s_1} = \omega_{s_1} \,|\, X_s = \omega_s, s \neq s_1) \\
&\qquad P_{k_0,k-1}(X_s = \omega_s, s \neq s_1) - \pi_{T(k,C)}(X_s = \omega_s, s \in \mathscr{S})| \\
&\quad = \sum_{(\omega_{s_2},\dots,\omega_{s_N})} \Bigg(\sum_{\omega_{s_1} \in \Lambda} \pi_{T(k,C)}(X_{s_1} = \omega_{s_1} \,|\, X_s = \omega_s, s \neq s_1) \,| \\
&\qquad P_{k_0,k-1}(X_s = \omega_s, s \neq s_1) - \pi_{T(k,C)}(X_s = \omega_s, s \neq s_1)| \Bigg) \\
&\quad = \sum_{(\omega_{s_2},\dots,\omega_{s_N})} |P_{k_0,k-1}(X_s = \omega_s, s \neq s_1) \\
&\qquad - \pi_{T(k,C)}(X_s = \omega_s, s \neq s_1)| \\
&\quad = \sum_{(\omega_{s_2},\dots,\omega_{s_N})} \Bigg| \sum_{\omega_{s_1}} (P_{k_0,k-1}(X_s = \omega_s, s \in \mathscr{S}) \\
&\qquad - \pi_{T(k,C)}(X_s = \omega_s, s \in \mathscr{S})) \Bigg| \\
&\quad \le \sum_{(\omega_{s_1},\dots,\omega_{s_N})} |P_{k_0,k-1}(X_s = \omega_s, s \in \mathscr{S}) \\
&\qquad - \pi_{T(k,C)}(X_s = \omega_s, s \in \mathscr{S})| \\
&\quad = \|P_{k_0,k-1} - \pi_{T(k,C)}\|.
\end{aligned}$$

Second, we prove that $\pi_{T(k,C)}$ converges to π_0 (the uniform distribution on Ω_{opt}):

$$\lim_{k\to\infty} \|\pi_0 - \pi_{T(k,C)}\| = 0.$$

To see this, let $|\Omega_{\text{opt}}|$ be the number of globally optimal configurations. Then

$$\begin{aligned}
\lim_{k\to\infty} \pi_{T(k,C)}(\omega) &= \lim_{k\to\infty} \frac{\exp(-U(\omega) \oslash T(k,C))}{\sum_{\omega' \in \Omega_{\text{opt}}} \exp(-U(\omega') \oslash T(k,C)) + \sum_{\omega' \notin \Omega_{\text{opt}}} \exp(-U(\omega') \oslash T(k,C))} \\
&= \lim_{k\to\infty} \frac{\exp(-(U(\omega) - U^{\text{inf}}) \oslash T(k,C))}{|\Omega_{\text{opt}}| + \sum_{\omega' \notin \Omega_{\text{opt}}} \exp(-(U(\omega) - U^{\text{inf}}) \oslash T(k,C))} = \begin{cases} 0 & \omega \notin \Omega_{\text{opt}} \\ \dfrac{1}{|\Omega_{\text{opt}}|} & \omega \in \Omega_{\text{opt}}. \end{cases}
\end{aligned} \tag{71}$$

The above equation is true if $(U(\omega) - U^{\text{inf}}) \oslash T(k,C) \ge 0$. Let us rewrite this inequality as

$$\sum_{C \in \mathscr{C}} \frac{V_C(\omega) - V_C(\omega')}{T(k,C)} \ge 0, \tag{72}$$

where ω' is any globally optimal configuration (i.e., $\omega' \in \Omega_{\text{opt}}$). While $V_C(\omega) - V_C(\omega')$ may be negative, $U(\omega) - U^{\text{inf}}$ is always positive or zero. We denote by $\Sigma(\omega)$ the energy difference in Eq. (72) without the temperature. Obviously, it is nonnegative:

$$\Sigma(\omega) = \sum_{C \in \mathscr{C}} V_C(\omega) - V_C(\omega') = U(\omega) - U^{\text{inf}} \ge 0.$$

Then let us decompose $\Sigma(\omega)$ according to Eq. (47):

$$\Sigma(\omega) = \Sigma^{+}(\omega, \omega') + \Sigma^{-}(\omega, \omega'),$$

from which

$$\Sigma^{+}(\omega, \omega') = \Sigma(\omega) - \Sigma^{-}(\omega, \omega').$$

Now, we consider Eq. (72):

$$\begin{aligned}
&\sum_{C \in \mathscr{C}} \frac{V_C(\omega) - V_C(\omega')}{T(k,C)} \\
&\quad = \Sigma^{-}(\omega, \omega') \oslash T(k,C) + \Sigma^{+}(\omega, \omega') \oslash T(k,C) \\
&\quad \ge \Sigma^{-}(\omega, \omega')/T_k^{\text{inf}} + \Sigma^{+}(\omega, \omega')/T_k^{\text{sup}} \\
&\quad = \frac{\Sigma^{-}(\omega, \omega') \cdot T_k^{\text{sup}} + \Sigma^{+}(\omega, \omega') \cdot T_k^{\text{inf}}}{T_k^{\text{inf}} T_k^{\text{sup}}} \ge 0.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\Sigma^{-}(\omega, \omega') \cdot T_k^{\text{sup}} + \Sigma^{+}(\omega, \omega') \cdot T_k^{\text{inf}} \\
&\quad = \Sigma^{-}(\omega, \omega') \cdot T_k^{\text{sup}} + (\Sigma(\omega) - \Sigma^{-}(\omega, \omega'))T_k^{\text{inf}}.
\end{aligned}$$

Therefore,

$$\Sigma^{-}(\omega, \omega')(T_k^{\sup} - T_k^{\inf}) - \Sigma(\omega) \cdot T_k^{\inf} \geq 0.$$

Dividing by $\Sigma^{-}(\omega, \omega')$ which is negative, we get

$$T_k^{\sup} - T_k^{\inf} \leq \frac{\Sigma(\omega)}{|\Sigma^{-}(\omega, \omega')|} T_k^{\inf},$$

which is true due to condition 3 of the theorem. Finally, we can prove that

$$\sum_{k=1}^{\infty} \|\pi_{T(k,C)} - \pi_{T(k+1,C)}\| < \infty, \tag{73}$$

since

$$\sum_{k=1}^{\infty} \|\pi_{T(k,C)} - \pi_{T(k+1,C)}\| = \sum_{\omega} \sum_{k=1}^{\infty} |\pi_{T(k,C)}(\omega) - \pi_{T(k+1,C)}(\omega)|,$$

and since

$$\forall\omega\colon \pi_{T(k,C)}(\omega) \to \pi_0(\omega),$$

it is enough to show that $\pi_T(\omega)$ is monotonous for every ω. However, it is clear from Eq. (71) that

- if $\omega \notin \Omega_{\text{opt}}$ then $\pi_T(\omega)$ is strictly increasing for $0 < T \leq \varepsilon$ for some sufficiently small ε,
- if $\omega \in \Omega_{\text{opt}}$ then $\pi_T(\omega)$ is strictly decreasing for all $T > 0$.

Fix $k > k_0 \geq 0$. From Eq. (70) and Eq. (73), we obtain

$$\begin{aligned}
\|P_{k_0,k} - \pi_0\| &\leq \|P_{k_0,k} - \pi_{T(k,C)}\| + \|\pi_{T(k,C)} - \pi_0\| \\
&\leq \|P_{k_0,k-1} - \pi_{T(k,C)}\| + \|\pi_{T(k,C)} - \pi_0\| \text{ by Eq. (70)} \\
&\leq \|P_{k_0,k-1} - \pi_{T(k-1,C)}\| + \|\pi_{T(k-1,C)} \\
&\quad - \pi_{T(k,C)}\| + \|\pi_{T(k,C)} - \pi_0\| \\
&\leq \|P_{k_0,k-2} - \pi_{T(k-2,C)}\| + \|\pi_{T(k-2,C)} - \pi_{T(k-1,C)}\| \\
&\quad + \|\pi_{T(k-1,C)} - \pi_{T(k,C)}\| + \|\pi_{T(k,C)} - \pi_0\| \\
&\leq \cdots \leq \|P_{k_0,k_0} - \pi_{T(k_0,C)}\| + \sum_{l=k_0}^{k-1} \|\pi_{T(l,C)} - \pi_{T(l+1,C)}\| \\
&\quad + \|\pi_{T(k,C)} - \pi_0\|.
\end{aligned}$$

On the other hand,

$$P_{k_0,k_0} = \pi_0$$

and

$$\lim_{k\to\infty} \|\pi_{T(k,C)} - \pi_0\| = 0.$$

Then we have

$$\begin{aligned}
\overline{\lim_{k_0\to\infty}} \sup_{k\geq k_0} \|P_{k_0,k} - \pi_0\| &\leq \overline{\lim_{k_0\to\infty}} \sup_{k>k_0} \sum_{l=k_0}^{k-1} \|\pi_{T(l,C)} - \pi_{T(l+1,C)}\| \\
&= \overline{\lim_{k_0\to\infty}} \sum_{l=k_0}^{\infty} \|\pi_{T(l,C)} - \pi_{T(l+1,C)}\| = 0.
\end{aligned}$$

The last term is 0 by (73) which completes the proof of Lemma A.1. Q.E.D.

Theorem 4.1 (Multitemperature Annealing). *Assume that there exists an integer $\kappa \geq N$ such that for every $k = 0, 1, 2, \ldots, \mathcal{S} \subseteq \{n_{k+1}, n_{k+2}, \ldots, n_{k+\kappa}\}$. For all $C \in \mathcal{C}$, let $T(k, C)$ be any decreasing sequence of temperatures in k for which*

1. $\lim_{k\to\infty} T(k, C) = 0$. *Let us denote respectively by $T_k^{\inf}$ and $T_k^{\sup}$ the maximum and minimum of the temperature function at k ($\forall C \in \mathcal{C}$: $T_k^{\inf} \leq T(k, C) \leq T_k^{\sup}$).*
2. *For all $k \geq k_0$, for some integer $k_0 \geq 2$: $T_k^{\inf} \geq N\Sigma_\Delta^{+}/\ln(k)$.*
3. *If $\Sigma^{-}(\omega, \omega') \neq 0$ for some $\omega \in \Omega \backslash \Omega_{\text{opt}}$, $\omega' \in \Omega_{\text{opt}}$, then a further condition must be imposed: For all k, $T_k^{\sup} - T_k^{\inf}/T_k^{\inf} \leq R$ with*

$$R = \min_{\substack{\omega\in\Omega\backslash\Omega_{\text{opt}} \\ \omega'\in\Omega_{\text{opt}} \\ \Sigma^{-}(\omega,\omega')\neq 0.}} \frac{U(\omega) - U^{\inf}}{|\Sigma^{-}(\omega,\omega')|}$$

Then for any starting configuration $\eta \in \Omega$ and for every $\omega \in \Omega$,

$$\lim_{k\to\infty} P(X(k) = \omega \,|\, X(0) = \eta) = \pi_0(\omega). \tag{74}$$

Proof. Using the above-mentioned lemmas, we can easily prove the annealing theorem:

$$\begin{aligned}
&\overline{\lim_{k\to\infty}} \|P(X(k) = \cdot|X(0) = \eta) - \pi_0\| \\
&\quad = \overline{\lim_{k_0\to\infty}} \overline{\lim_{\substack{k\to\infty \\ k\geq k_0}}} \Big\| \sum_{\eta'} P(k, \cdot|k_0, \eta')P(k_0, \eta'|0, \eta) - \pi_0\Big\| \\
&\quad \leq \overline{\lim_{k_0\to\infty}} \overline{\lim_{\substack{k\to\infty \\ k\geq k_0}}} \Big\| \sum_{\eta'} P(k, \cdot|k_0, \eta')P(k_0, \eta'|0, \eta) \\
&\qquad - P(k, \cdot|k_0, \pi_0)\Big\| + \overline{\lim_{k_0\to\infty}} \overline{\lim_{\substack{k\to\infty \\ k\geq k_0}}} \|P(k, \cdot|k_0, \pi_0) - \pi_0\|.
\end{aligned}$$

The last term is 0 by Lemma A.2. Moreover, $P(k_0, \cdot|0, \eta)$ and π_0 have total mass 1; thus,

$$\begin{aligned}&\left\|\sum_{\eta'} P(k, \cdot|k_0, \eta')P(k_0, \eta'|0, \eta) - P(k, \cdot|k_0, \pi_0)\right\|\\ &\quad = \sum_{\omega} \sup_{\eta''} \Big|\sum_{\eta'} (P(k, \omega|k_0, \eta')\\ &\qquad - P(k, \omega|k_0, \eta''))(P(k_0, \eta'|0, \eta) - \pi_0(\eta'))\Big|\\ &\quad \leq 2 \sum_{\omega} \sup_{\eta', \eta''} |P(k, \omega|k_0, \eta') - P(k, \omega|k_0, \eta'')|.\end{aligned}$$

Finally,

$$\begin{aligned}&\overline{\lim_{k\to\infty}} \|P(X(k) = \cdot|X(0) = \eta) - \pi_0\|\\ &\qquad \leq 2 \sum_{\omega} \overline{\lim_{k_0\to\infty}}\; \overline{\lim_{\substack{k\to\infty\\ k\geq k_0}}} \sup_{\eta', \eta''} |P(k, \omega|k_0, \eta')\\ &\qquad\quad - P(k, \omega|k_0, \eta'')| = 0.\end{aligned}$$

The last term is 0 by Lemma A.1 which completes the proof of the annealing theorem. Q.E.D.

REFERENCES

1. R. Azencott, Markov fields and image analysis, in *Proceedings AFCET, Antibes, 1987.*
2. R. Azencott, *Parallel Simulated Annealing: An Overview of Basic Techniques,* pp. 37–46. Wiley, New York, 1992.
3. J. Besag, On the statistical analysis of dirty pictures, *J. R. Statist. Soc. B,* 1986.
4. J. E. Besag, Spatial interaction and the statistical analysis of lattice systems (with discussion), *J. R. Statist. Soc. B* **36,** 1974, 192–236.
5. A. Blake and A. Zisserman, *Visual Reconstruction,* MIT Press, Cambridge, MA, 1987.
6. C. Bouman and B. Liu, Multiple resolution segmentation of texture images, *IEEE Trans. Pattern Anal. Mach. Intell.* **13,** 1991, 99–113.
7. C. A. Bouman and M. Shapiro, A multiscale random field model for Bayesian image segmentation, *IEEE Trans. Image Process.* **3**(2), 1994, 162–177.
8. B. Chalmond, Image restoration using an estimated markov model, *Signal Process.* **15,** 1988, 115–129.
9. D. Geiger and F. Girosi, Parallel and deterministic algorithms for MRFs: Surface reconstruction and integration, in *Proceedings ECCV90, Antibes, France, 1990,* pp. 89–98.
10. D. Geiger and J. Kogler, Scaling images and image features via the renormalization group, in *Proceedings IEEE CVPR '93, New York, June 1993.*
11. S. Geman and D. Geman, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Trans. Pattern Anal. Mach. Intell.* **6,** 1984, 721–741.
12. S. Geman and C. Graffigne, Markov random fields image models and their application to computer vision, in *Proceedings ICM '86* (A. M. Gleason, Ed.), Amer. Math. Soc., Providence, 1987.
13. B. Gidas, A renormalization group approach to image processing problems, *IEEE Trans. Pattern Anal. Mach. Intell.* **11**(2), 1989, 164–180.
14. F. Heitz, P. Pérez, and P. Bouthemy, Multiscale minimization of global energy functions in some visual recovery problems. *CVGIP: Image Understanding* **59**(1), 1994, 125–134.
15. F. Heitz, P. Pérez, E. Memin, and P. Bouthemy, Parallel visual motion analysis using multiscale Markov random fields, in *Proceedings of Workshop on Motion, Princeton, Oct. 1991.*
16. W. D. Hillis, *The Connection Machine,* MIT Press, Cambridge, MA, 1985.
17. M. Hurn and C. Jennison, A study of simulated annealing and a revised cascade algorithm for image reconstruction, Technical Report 93:04, University of Bath, Apr. 1993.
18. F. C. Jeng and J. M. Woods, Compound Gauss–Markov random fields for image estimation, *IEEE Trans. Acoust. Speech Signal Process.* **39,** 1991, 638–697.
19. Z. Kato, M. Berthod, and J. Zerubia, Multiscale Markov random field models for parallel image classification, in *Proceedings ICCV, Berlin, May 1993.*
20. Z. Kato, M. Berthod, and J. Zerubia, Parallel image classification using multiscale Markov random fields, in *Proceedings ICASSP, Minneapolis, Apr. 1993.*
21. Z. Kato, J. Zerubia, and M. Berthod, Bayesian image classification using Markov random fields, in *Proceedings Maxent Workshop, Paris, July 1992.*
22. Z. Kato, J. Zerubia, and M. Berthod, Satellite image classification using a modified Metropolis dynamics, in *Proceedings ICASSP, San Francisco, CA, Mar. 1992.*
23. P. V. Laarhoven and E. Aarts, *Simulated Annealing: Theory and Applications,* Reidel, Dordrecht, Holland, 1987.
24. N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, Equation of state calculations by fast computing machines, *J. Chem. Phys.* **21,** 1953, 1987–1092.
25. G. K. Nicholls and M. Petrou, A generalisation of renormalisation group methods for multiresolution image analysis, in *Proceedings ICPR '92, 1992,* pp. 567–570.
26. P. Pérez, Champs markoviens et analyse multirésolution de l'image: Application à l'analyse du mouvement, Ph.D. thesis, Université de Rennes I, 1993.
27. J. Zerubia and R. Chellappa, Mean field annealing using compound Gauss–Markov random fields for edge detection and image estimation, *IEEE Trans. Neural Networks* **8**(4), 1993, 703–709.