An adaptive simulated annealing cooling schedule for object detection in images

Mathias Ortner, Xavier Descombes, Josiane Zerubia

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

An adaptive simulated annealing cooling schedule for object detection in images

Mathias Ortner — Xavier Descombes — Josiane Zerubia

N° 6336

October 2007

INRIA
SOPHIA ANTIPOLIS

An adaptive simulated annealing cooling schedule for object detection in images

Mathias Ortner*, Xavier Descombes*, Josiane Zerubia *

Thème COG — Systèmes cognitifs
Projets Ariana

Rapport de recherche n° 6336 — October 2007 — 27 pages

Abstract: In our image processing applications, we use a simulated annealing procedure to find configurations of geometric shapes that fit the best an image. This type of algorithm allows finding one of the global minima of an arbitrary function provided that the cooling schedule is logarithmic with the time. Since this type of cooling schedules is very slow, geometrical cooling schemes are used in practice. Geometrical schemes are however subject to some disadvantages that we discuss in this report. To overcome these disadvantages, we propose an adaptive cooling scheme. This heuristic is based on the analysis of the cooling scheme behavior in practice. In particular, we observe the presence of critical temperatures. To deal with these critical temperatures, we propose a cooling scheme that decelerates when such a temperature is detected, and accelerates otherwise. We present results on a real problem taken from our image processing applications.

Key-words: Image processing, shape extraction, spatial point process, simulated annealing, adaptive cooling schedule.

* Ariana - Joint research group CNRS / INRIA / UNSA - INRIA, 2004 route des Lucioles, BP 93, Sophia Antipolis, France.
Tel : +33 492 387 857, Fax : +33 492 387 643, Email : Firstname.Lastname@sophia.inria.fr

Recuit simulé adaptatif pour la détection d'objets dans les images

Résumé : Dans nos applications, nous utilisons des techniques de recuit-simulé pour trouver la configuration de formes géométriques qui décrit le mieux une image. Ce type d'algorithme permet de trouver l'un des minimum globaux d'une fonction, dès lors que le shéma de décroissance de la température est logarithmique. La lenteur rédhibitoire de ce type de shéma impose en pratique l'utilisation d'une décroissance géométrique. Le shéma de décroissance géométrique présente toutefois des défauts que nous décrivons dans ce rapport. Pour pallier ces inconvénients, nous proposons un shéma de décroissance adaptatif. Ce shéma est construit à partir d'observations expérimentales. En particulier, nous observons la présence de températures critiques correspondant à des moments cruciaux du processus d'optimisation. Le shéma que nous proposons repose sur la détection de telles températures. Le shéma de refroidissement proposé ralentit lorsqu'une température critique est détectée, et accél'ere autrement. Nous présentons des résultats sur des problèmes réels tirés de nos applications en traitement d'images.

Mots-clés : Traitement d'images, extraction de formes, processus ponctuels spatiaux, recuit-simulé, shéma de décroissance adaptatif.

Contents

1 Introduction

In our image processing applications, we focus on the task of automatically detecting objects in real images. In [17, 16] we propose different models for detecting buildings from Digital Elevation Models (DEMs). We consider configurations of random geometric shapes (e.g. of rectangles, segments or both.) We use a spatial point process framework which has the following advantages. First it allows considering an unknown number of objects, and therefore does not require a precise knowledge on the number of objects to be detected. Moreover, spatial point processes can be specified through a density. The notion of a *probability density function* is essential for modeling tasks. The nice thing about spatial point process models is that the distribution of a Poisson point process can play the analog role to Lebesgue measure for real random variables. This notion of spatial point process density for instance enables the inclusion of a *prior knowledge* within the extraction procedure through a Bayesian model. Considering exponential families, we focus on sufficient statistics that account for the spatial

repartition of objects in the scene. Examples of such statistics are the number of aligned rectangles and the number of connected segments.

Alternatively, these models can be written under a Gibbs form, consisting in the *partition function* (normalizing constant) along with the system's energy function. This energy is actually divided into two terms. First, *the data term* quantifies the quality of a configuration of objects with respect to the data. Second, *the internal field* acts on the repartition of objects and allows introducing a basic knowledge on patterns of interest through the definition of interactions. These two parts play the role of a Likelihood and a prior model in a Bayesian framework.

As a result of this modeling step, the energy is usually highly complex and exhibits numerous local minima. We consequently use a *simulated annealing* procedure based on a Hastings Metropolis scheme. The first advantage of this algorithm is that it can be applied to cases where the normalizing constant is not known. The second advantage is that the procedure provides a global minimum of the energy. However, the latter result only holds for specific cases that are not met in practice.

The simulated annealing procedure has been extensively used over the years. Originally proposed by S. Kirkpatrick and al in [11], the method quickly found a great audience outside the optimization communnity and is one of the few optimization methods widely known in very different fields. General overviews and discussion can be found in [24, 5, 9, 2, 7]. Originally proposed through an analogy with thermodynamics, the method has been subject to deep analysis of its convergence properties (see [1,12,13,27,14]). In image processing, the method has been recurrently associated with Markov Random Fields (see [28] and references herein). The basic idea is to drive an exploring process towards one of the points of global minimima. The process is usually compared (see [11]) with the simulation of a system of particles that goes from high temperatures to a null temperature. Cooling the system slowly enough result in freezing the particles in the state achieving the global minimum of energy. However, the cooling schedule needs to be logarithmic, meaning that any arbitrary temperature is reached in an exponential time, making the procedure tremendously slow. In practice, geometric cooling schedules are therefore employed. The issue of an optimal cooling sheme has been the subject of many works. In [21] toy examples are used to derive analytical results that are compared to the usual assertions. Other works on cooling schedules include [4, 10, 22].

In our case, these geometrical cooling schemes face some disadvantages that we describe in this report. In particular the presence of critical temperatures considerably degrades the result quality. We present an adaptive cooling scheme based on experimental observations that allows slowing down the cooling scheme around such critical temperatures.

In Section 2 we describe briefly our image processing framework. In Section 3 we present the simulated annealing procedure, and different usual cooling schedules. In Section 4 we expose experimental observations that lead us to propose a new adaptive scheme in Section 5. We present results on a real problem taken from [16] in Section 6 and detail the possible future work in Section 7.

2 Image, shapes, point processes and energy

We model a 2D image as a *continuous* bounded set $K = [0, X_{1_{max}}] \times [0, X_{2_{max}}]$, and note $x = (c_1, c_2)$ a point of K.

2.1 Random configuration of points

A *configuration of points* $\mathbf{x}$ (noted in bold) is a unordered set of points in K

$$\mathbf{x} = \{x_1, \dots, x_{n(\mathbf{x})}\}, \qquad x_i \in K, \tag{1}$$

where $n(\mathbf{x}) = \text{card}(\mathbf{x})$ denotes the number of points in the configuration. We note $\mathcal{C}$ the set of all possible finite configurations. Let consider a mapping from an abstract probability space $(\Omega, \mathcal{A}, \mathbf{P})$ to the set of configurations $\mathcal{C}$. Due to the finiteness of the considered configurations along with the boundedness of K, the σ-algebra associated with $\mathcal{C}$ is well defined (see [26] for details.)

A *point process* $\mathbf{X}$ of points in K is such a measurable mapping

$$\forall \omega \in \Omega, \quad \mathbf{X}(\omega) = \{x_1, \dots, x_n, \dots\} \;\; x_i \in K. \tag{2}$$

Accordingly, a point process is a random variable whose realizations are random configurations of points.

2.2 Poisson point process

The most random point process (in the entropy sense) is the *Poisson point process*. Let $\nu(.)$ be a positive measure on K. A Poisson point process $\mathbf{X}$ with intensity $\nu(.)$ verifies the following properties

- for every Borel set $A \subset K$, the random variable $N_{\mathbb{X}}(A)$, giving for the number of points of $\mathbf{X}$ falling in the set A, follows a discrete Poisson distribution with mean $\nu(A)$

$$\mathbf{P}(N_{\mathbb{X}}(A) = n) = e^{-\nu(A)} \frac{\nu(A)^n}{n!}, \tag{3}$$

- and for every finite sequence of non intersecting Borelian sets $B_1, \ldots, B_p$ the corresponding random variables $N_{\mathbb{X}}(B_1), \ldots, N_{\mathbb{X}}(B_p)$ are independent.

Poisson point processes are useful in our setup due to their analog role to Lebesgue measures on $\mathbb{R}^d$. As we detail it later, it is indeed possible to define point processes by specifying their density with respect to the distribution of a reference Poisson point process.

2.3 Marked point process

The configurations of points described so far only include simple points of $\mathbb{R}^2$. To describe random configurations of geometrical objects, random marks are added to each point.

For instance, let consider the following mark set

$$M^r = [-\frac{\pi}{2}, \frac{\pi}{2}] \times [L^r_{\min}, L^r_{\max}] \times [l^r_{\min}, l^r_{\max}]. \tag{4}$$

Noting by x elements of $S_r = K \times M_r$, we consider the following parameterization describing rectangles

$$x \in S_r, \quad x = (c_1(x), c_2(x), \theta(x), L(x), l(x)), \tag{5}$$

where $(c_1(x), c_2(x))$, $\theta(x)$, $L(x)$ and $l(x)$ correspond respectively to the center position, the orientation, the length and the width of the rectangle x. A marked point process $\mathbb{X}$ of rectangles is a point process on $S_r = K \times M_r$ [1]. Note that in [16] we also consider a process of segments.

2.4 Density of a spatial point process

An attractive feature of spatial point processes is the possibility of defining a point process distribution by its *probability density function* (pdf). A Poisson point process can indeed play the analog role to Lebesgue measure on $\mathbb{R}^d$.

Consider the distribution $\mu(.)$ of a Poisson point process defined by its non atomic intensity measure $\nu(.)$ and a mapping $h(.)$ from the space of configurations of points $\mathcal{C}$ to $[0, \infty[$. We consider the function $Z(\mu, h)$ defined as

$$Z = \int_{\mathcal{C}} h(\mathbf{x}) d\mu(\mathbf{x}). \tag{6}$$

If $Z < \infty$, the function $Z^{-1}h(\mathbf{x})$ can be seen as the density of a point process $\mathbb{X}$ with respect to the reference Poisson process (see [26]).

For instance, assume that

$$h(\mathbf{x}) = \prod_{i=1}^{n(\mathbf{x})} \beta(x_i) \tag{7}$$

where $\beta(.)$ is an *intensity* function from S to $]0, \infty[$. A point process $\mathbb{X}$ defined by this density turns to be a Poisson point process with intensity

$$\nu'(A) = \int_A \beta(u) d\nu(u). \tag{8}$$

In this simple case, the probability density function $Z^{-1}h(.)$ allows a change of intensity measure. This example actually belongs to the more general class of *exponential families*. Let $t(.)$ be a mapping from $\mathcal{C}$ to $\mathbb{R}^k$. It is

[1] There is actually a further requirement that the restriction of $\mathbb{X}$ to K, noted $\mathbb{X}_{|K}$, should also be a point process on K. In our case, this technical condition on the measurability of the mapping $\mathbb{X}$ is satisfied since the sets K and S_r are bounded, see [26] for details.

possible to describe a class of point process densities by using a parameter $\theta \in \mathbb{R}^k$ together with the scalar product $< ., . >$

$$h(\mathbf{x}) = e^{-<\theta, t(\mathbf{x})>} \tag{9}$$

Of course, the density is well defined if and only if $Z(\theta, \mu) < \infty$. In this work we introduce a density where points are not independent but are correlated by means of interaction energies. In our applications we usually consider the density from the energy point of view:

$$U(\mathbf{x}) = - < \theta, t(\mathbf{x}) > . \tag{10}$$

2.5 Estimator and MCMC

In [15] we presented an MCMC algorithm generating samples of a point process $\mathbf{X}$ defined by an unnormalized density $h(.)$ along with a reference Poisson point process distribution. The obtained algorithm produces a Markov Chain $(X_t)_{t \geq 0}$ ergodically converging to the distribution of $\mathbf{X}$.

The procedure enables the computation of Monte Carlo values. Another possibility is to use the sampler within a *simulated annealing* framework providing a global maximum of the density $h(.)$ as described in [25]. The estimator obtained is consequently the maximum density estimator

$$\hat{\mathbf{x}} = \text{Argmax } h(.) \tag{11}$$

We detail these notions in Section 3.

2.6 Energy

In our applications [17, 16] we define models through the energy of a configuration of points. These models can be alternatively seen as exponential families, thereby allowing parameter estimation procedures (see [6]). The energy is defined over the set of configurations of objects

$$\begin{array}{rccc} U: & \mathcal{C} & \rightarrow & \bar{\mathbb{R}} \\ & \mathbf{x} & \rightarrow & U(\mathbf{x}) \end{array} \tag{12}$$

Basically the energy models we consider are made of two parts

$$U(\mathbf{x}) = U_{data}(\mathbf{x}) + \rho U_{reg}(\mathbf{x}), \qquad \rho > 0. \tag{13}$$

The data term $U_{data}(\mathbf{x})$ quantifies the relevance of the configuration $\mathbf{x}$ with respect to the data, while $U_{reg}(\mathbf{x})$ acts on the geometric pattern of objects. For instance, in [17] we use a regularization term that favors alignments between rectangles.

2.7 Examples of result

We present here some results taken from [16]. We consider a process of rectangles and segments, and a prior model that favors a paving behavior of the rectangles as well as connection interactions between segments. Two samples of the prior model are presented in Figure 1. The data term is designed such that the segments are attracted by the discontinuities of the image while the rectangles by the homogenous areas. We also consider a term that favor interactions between both kind of geometric shapes, in order to make both informations (discontinuity and homogeneity) fit one each other. Figure 2 presents an extraction result on a Digital Elevation Model provided by the French National Geographic Institute (IGN).

3 Simulated annealing

In this section, we describe the simulated annealing procedure. This type of algorithm has been widely used in image processing (see [28], for instance). We first present the greedy algorithm which is the simplest stochastic optimization algorithm, and expose the simulated annealing that can be intuitively seen as an extension of the former.

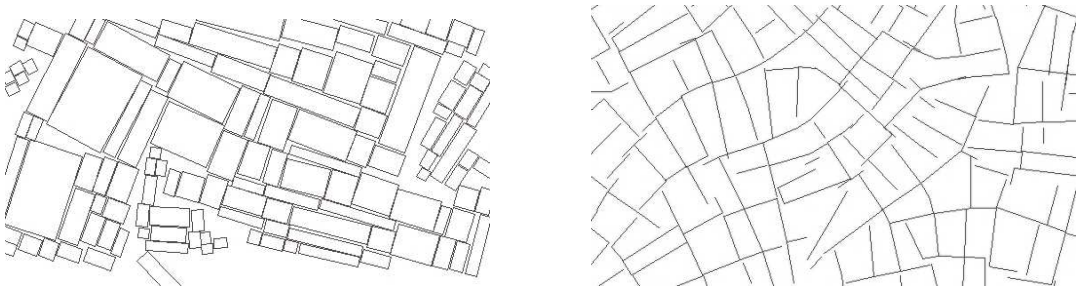

Figure 1: Simulation result showing the influence of the internal field on the process of rectangles and segments.

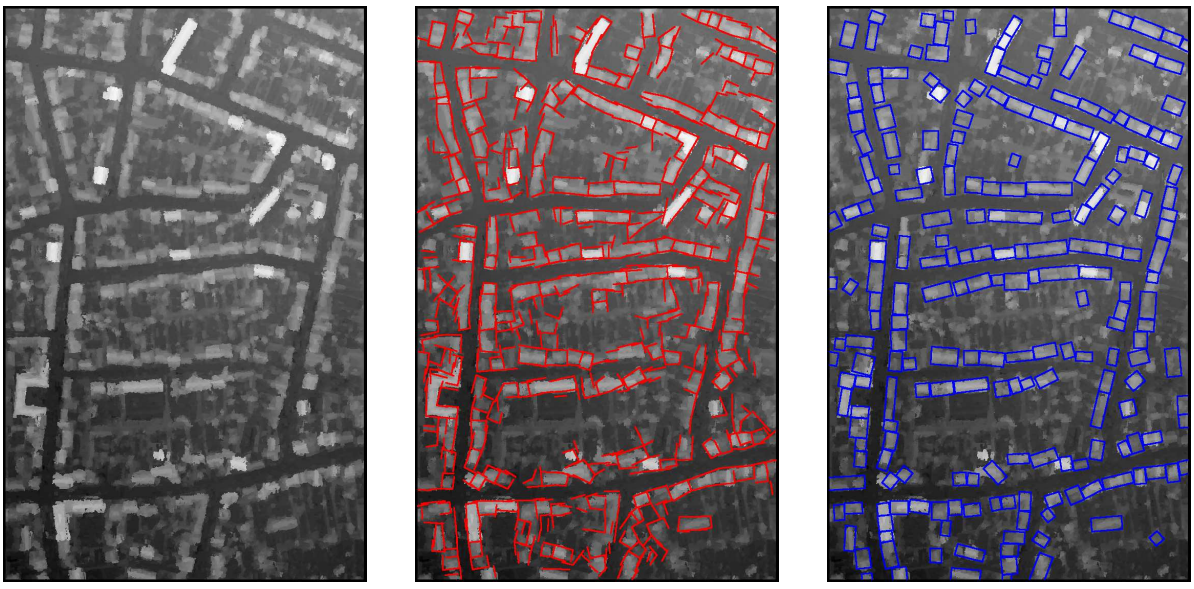

Figure 2: From left to right. Digital Elevation Model of a part of Rennes, France (©IGN) that has been obtained by aerial stereovision, segment extraction result, and rectangle extraction result.

3.1 Greedy algorithm

The simplest way of minimizing $U(.)$ is to use a greedy algorithm. This procedure starts from an arbitrary initial state X_0. At each time step t, the current configuration $X_t = \mathbf{x}$ is randomly modified, resulting in a new configuration $\mathbf{y}$. The perturbation induces an energy variation

$$\Delta U = U(\mathbf{y}) - U(\mathbf{x}).$$

The greedy algorithm accepts to replace $\mathbf{x}$ by $\mathbf{y}$ only if $\Delta U \leq 0$, i.e. only if the perturbation has improved the current configuration. This rather simple procedure is then iterated. Addition, deletion, translation, rotation or modification of a randomly chosen object are examples of usual transformations. The major and obvious drawback of this algorithm is that it provides a local minimum that depends on the initial configuration as well as on the set of possible transformations.

3.2 Hastings Metropolis and simulated annealing

3.2.1 Generic Structure

Suppose we consider a point process $\mathbf{X}$ defined by its energy $U(.)$. Through the Gibbs relation, this energy leads to a density h known up to a normalizing constant. This density together with the distribution $\mu(.)$ of the reference Poisson point process defines the distribution $\pi(.)$ of $\mathbf{X}$.

The Markov chain $(X_t)_{t\geq 0}$ is defined by a starting point $X_0 = \{\emptyset\}$ and a Markovian transition kernel $P(\digamma, .)$ corresponding to the conditional distribution of $X_{t+1}|X_t = \mathbf{x}$. It results in a Markov chain $(X_t)_{t\geq 0}$ on the space of finite configurations of points $\mathcal{C}$.

Of course, $P(.,.)$ is designed in order to make the Markov Chain converge towards the desired distribution.

$$\|P^n(\{\emptyset\}, .) - \pi(.)\|_{TV} \to 0 \tag{14}$$

where $\|.\|_{TV}$ notes the Total Variation norm (TV).

The Markov chain generated by the following algorithm satisfies this property. We actually have more accurate results, since we know that we can start from any configuration (Harris recurrence) and that the total variation tends to zero geometrically (geometric ergodicity), as detailed in [15].

3.2.2 Algorithm

The algorithm is based on a mixture of perturbation kernels $Q(.,.) = \sum_m p_m Q_m(.,.)$ where $\sum p_m = 1$ and $\int Q_m(\mathbf{x}, \mathbf{x}')\mu(d\mathbf{x}) = 1$. The algorithm iterates the following steps. We fix the current state X_t as $X_t = \mathbf{x} = \{z_1, \ldots, z_n\}$.

[1] Choose one of the proposition kernels $Q_m(.,.)$ with probability $p_m(\mathbf{x})$ and

[2] sample $\mathbf{x}'$ according to the chosen kernel $\mathbf{x}' \sim Q_m(\mathbf{x}, .)$.

[3] Compute the Green's ratio $R_m(\mathbf{x}, \mathbf{x}')$, function of the selected kernel Q_m, the original state $\mathbf{x}$ and the proposed new state $\mathbf{x}'$. The ratio R_m is derived to make the Markov chain converge towards the desired distribution.

[4] The proposition is accepted $X_{t+1} = \mathbf{x}'$ with a probability $\alpha_m(\mathbf{x}, \mathbf{x}') = min(R_m(\mathbf{x}, \mathbf{x}'), 1)$ and rejected otherwise.

3.2.3 Perturbation kernels

The efficiency of the algorithm highly depends on the variety of possible transformations $Q_m(\mathbf{x}, .)$.

Birth or death. This kind of perturbation first chooses with probability p_b and $p_d = 1 - pb$ whether a point should be removed (death) or added (birth) to the configuration. If death is chosen, the kernel selects randomly one point u in $\mathbf{x}$ and proposes $\mathbf{x}' = \mathbf{x} \setminus u$, while if birth is chosen, it generates a new point u according to the uniform measure $|.|/|S|$ and proposes $\mathbf{x}' = \mathbf{x} \cup u$. The birth or death kernel is necessary and sufficient to insure the convergence of the Markov chain towards the target distribution.

Non jumping transformations. Non jumping transformations are transformations that first select randomly a point u in the current configuration and then propose replacing this point by a perturbed version v, $\mathbf{x}' = \mathbf{x} \backslash u \cup v$. Translation, rotation or dilation are examples of non jumping perturbations.

Birth or death in a neighborhood. We introduced this kind of transformation in [15]. The idea is to propose the removal or addition of interacting pairs of points with respect to one of the attractive relations such as the connection in the case of segments or alignment in the case of rectangles.

Green ratio. With each of these proposition kernels a mapping $R_m(.,.)$ from $\mathcal{C} \times \mathcal{C}$ to $(0, \infty)$ is associated. This value, named *Green ratio*, depends on the target distribution π. See [15,16,17] for derivations of Green ratios in specific cases.

3.2.4 Simulated annealing

To find a minimizer of the energy $U(.)$ we use a simulated annealing framework. Instead of generating samples of $h(.)$, we simulate $h^{\frac{1}{T_t}}(.)$. The temperature parameter T_t tends to zero as t tends to ∞. Note that it is equivalent to the notation

$$f_t(\mathbf{x}) = Z_{T_t}^{-1} \exp\left(-\frac{U(\mathbf{x})}{T_t}\right). \tag{15}$$

If T_t decreases with a logarithmic rate, then X_t converges a.s. towards one of the global maximizers of $h(.)$. The temperature parameter constructs a sequence of probability measures that almost surely converges pointwise towards a set of Dirac measures putting with masses on the set of global minima. The logarithmic schedule gives Dobrushin conditions asserting the convergence of process towards this set of Dirac masses (see [25]).

3.3 Logarithmic cooling schedule

It is well known (see [8]) that if a logarithmic cooling schedule

$$T_t = \frac{C}{\log(1+t)}$$

is adopted, then the Chain converges towards a global minimum of the energy, provided that C is greater than the depth of the local minimum that is not a global minimum. It should be noted that in the case of simulated annealing applied to point process models a proof of convergence exists, based on a "birth or death" algorithm ([25]) but to our knowledge, such a proof does not exist in the case of a Metropolis-Hastings update type, although the extension should be straightforward.

3.4 Geometrical cooling schedule

The logarithmic schedule exhibits a major drawback: the time required to reach any temperature follows an exponential function. Since in practice we are limited by finite times, a common solution is to use a geometrical cooling schedule

$$T_t = T_0 A^t,$$

where $A < 1$ tunes the cooling speed.

Note that the quality of the proposition kernels is an important issue. As a consequence we design kernels such that the trajectory of the Markov chain is poorly correlated to insure a good exploration of the state space, see [16].

A very similar scheme is the piecewise constant geometrical scheme. The idea is to keep the temperature constant on steps and decrease it in precise instants. At first glance, this idea looks rather useless since it is equivalent to increase the constant A. However, this idea has a main advantage. The step on which the temperature is kept constant can be used to test whether the Markov Chain has converged or not, and thereby decrease the temperature accordingly. For instance, in [3], Brooks proposes decreasing the temperature only if the chain has converged. Of course, this idea requires the existence of a test on the chain convergence, which is not a simple problem (see [18]).

3.5 Adaptive schedules

In [4] the author presents a comparative study of different proposition kernels and cooling schedules that can be used for a set of academic simulated annealing problems (including the Ising model). An interesting formalism is proposed, representing the selection of an optimal cooling schedule as a control problem. In some simple cases it is possible to derive analytical solutions. Nevertheless, the complexity of our models make analytical results difficult to obtain. We therefore focus on more generic heuristics. We present here two interesting approaches that can be found in the litterature, although they failed in our case.

Reversibility. One of the ideas studied in [4], originally proposed in [23], is based on a slightly modified geometrical cooling scheme. Let consider some time intervals (steps) $t \in [n_i, n_{i+1}[$ indexed by i and the associated average energy

$$\langle U \rangle_i = \frac{1}{n_{i+1} - n_i + 1} \sum_{k=n_i}^{n_{i+1}} U(\mathbf{x}_k). \tag{16}$$

This average energy at step i can be seen as a statistic. A basic idea is to compare the measured average level to the previous one (namely $\langle U \rangle_{i-1}$) and decrease the temperature only if the new average energy is greater than previous one

$$T_{i+1} = \begin{cases} T_i & \text{if} \quad \langle U \rangle_{i+1} \leq \langle U \rangle_i \\ AT_i & \text{if} \quad \langle U \rangle_{i+1} > \langle U \rangle_i. \end{cases} \tag{17}$$

Basically, this scheme relies on the idea that the stability of the average energy is a good indicator of the chain convergence. This scheme is subject to two major problems. First the constant A needs to be adjusted properly, and second, the scheme is very slow in practice.

Constant thermodynamic speed. Another possibility is to design a scheme by extending the intuitive physical analogy of the simulated annealing procedure. One idea is to focus on the entropy of the system. We describe here a scheme detailed in [20, 19].

In addition to the average energy, let consider a second order statistic

$$\langle (\Delta U)^2 \rangle = \frac{1}{n_{i+1} - n_i + 1} \sum_{k=n_i}^{n_{i+1}} (U(\mathbf{x}_k) - \langle U \rangle_i)^2. \tag{18}$$

The variance can be linked to the thermic capacity

$$C(T) = \frac{\langle (\Delta U)^2 \rangle}{T^2}, \tag{19}$$

which turns to be the derivative of the system entropy

$$\frac{dS}{dT} = \frac{C}{T}.$$

As a consequence, the following scheme exhibits a constant entropy variation

$$T_{i+1} - T_i \propto \frac{T}{C} = \frac{T_i^3}{\langle U^2 \rangle_i - \langle U \rangle_i^2}.$$

The proportionality parameter needs to be carefully tuned. Loosely speaking, this scheme tends to limit the gap of temperature between two steps when the energy variation is important. In our case, this scheme did not work well: the choice of the constant appeared to be a major issue, especially in order to get finite computational times.

4 Experimental observations

We describe in this section some experimental observations. We use the model we presented in [16] and which we briefly described in Section 2.7.