Theoretically Grounded Acceleration Techniques for Simulated Annealing

Marc C. Robini

Abstract. Simulated annealing (SA) is a generic optimization method whose popularity stems from its simplicity and its global convergence properties; it emulates the physical process of annealing whereby a solid is heated and then cooled down to eventually reach a minimum energy configuration. Although successfully applied to many difficult problems, SA is widely reported to converge very slowly, and it is common practice to relax some of its convergence conditions as well as to allow extra freedom in its design. However, variations on the theme of annealing usually come without optimal convergence guarantees.

In this paper, we review the fundamentals of SA and we focus on acceleration techniques that come with a rigorous mathematical justification. We discuss the design of the candidate-solution generation mechanism, the issue of finite-time cooling, and the technique of acceleration by concave distortion of the objective function. We also investigate a recently introduced generalization of SA — stochastic continuation — which significantly increases the design flexibility by allowing the candidate-solution generation mechanism and the objective function to vary with temperature.

1 Introduction

1.1 Background

Simulated annealing (SA) is a generic method for combinatorial optimization that is quite popular because of its ease of implementation and its global convergence properties. The key feature of SA is to allow uphill moves (that is, moves that increase the value of the objective function) in order to escape local minima. By analogy with the

Marc C. Robini
CREATIS (CNRS UMR 5220; INSERM U1044), INSA-Lyon,
69621 Villeurbanne cedex, France
e-mail: marc.robini@creatis.insa-lyon.fr

I. Zelinka et al. (Eds.): Handbook of Optimization, ISRL 38, pp. 311–335.
springerlink.com © Springer-Verlag Berlin Heidelberg 2012

physical process of annealing in solids, uphill moves are accepted with some probability controlled by a temperature parameter that decreases monotonically to zero. As the temperature goes to zero, the invariant measure of the underlying Markov chain model concentrates on the global minima of the objective function, and we can expect that the process converges to a global minimum if the cooling is sufficiently slow. Early results [15, 16, 9] show that this is indeed the case if the temperature is inversely proportional to the logarithm of the iteration index. However, this theoretical advantage is counterbalanced by well-known practical disadvantages, namely, that SA converges very slowly and that the convergence assumptions severely limits design freedom.

Good SA algorithm design means careful selection of the cooling schedule—most successful applications of SA use exponential cooling, which is theoretically justified in [6]—and clever construction of the candidate-solution generation mechanism (we call it the *communication mechanism* for short). Nevertheless, many implementations of SA found in the literature use inappropriate cooling schedules and crude communication mechanisms, which usually translates to convergence to poor local minima and sensitivity to initialization. It is therefore not surprising that SA is often abandoned in favor of other (mainly deterministic) optimization methods. The truth is that carefully designed annealing algorithms produce very good results for a wide class of problems. Yet, standard SA is generally much slower than deterministic methods, and it is common practice to relax some of its convergence conditions as well as to allow extra freedom in its design at the expense of losing optimal convergence guarantees.

In this paper, we focus on acceleration techniques that come with a rigorous mathematical justification; these include (i) restriction of the state space, transformation of the state space, and relaxation, (ii) proper selection of the cooling schedule, (iii) concave distortion of the objective function, (iv) temperature dependence of the objective function, and (v) temperature dependence of the communication mechanism.

1.2 Overview

We start by reviewing the fundamentals of Metropolis-type SA on a finite state space, which is the most popular and the best understood class of annealing algorithms. Let U be a real-valued function to be minimized over a finite state space Ω; call it the *energy function*. Without going into details, a Metropolis-type SA algorithm with energy U is a Markov chain $(X_n)_{n\in\mathbb{N}}$ on Ω whose transitions are guided by a communication mechanism θ and controlled by a *cooling sequence* $(\beta_n)_{n\in\mathbb{N}^*}$. The communication mechanism is a Markov matrix on Ω that gives the probabilities of the possible moves for generating a candidate solution from the current solution, and the cooling sequence is a divergent sequence of inverse temperatures acting on the rate of acceptance of uphill moves. The transitions of $(X_n)_n$ are defined as follows: for any $(x,y)\in\Omega^2$ such that $x\neq y$,

$$\mathsf{P}(X_n = y \,|\, X_{n-1} = x) = \begin{cases} \theta(x,y) & \text{if } U(y) \leqslant U(x), \\ \theta(x,y)\exp(-\beta_n(U(y) - U(x))) & \text{if } U(y) > U(x). \end{cases} \quad (1)$$

Putting it simply, downhill moves are unconditionally accepted, whereas an uphill move from x to y is accepted with probability $\exp(-\beta_n(U(y) - U(x)))$ at iteration n. It is well known [16] that, under weak assumptions on θ, if $(\beta_n)_n$ increases slowly enough, then $(X_n)_n$ converges to the set of global minima of U in the sense that

$$\lim_{n\to+\infty} \mathsf{P}\big(U(X_n) > \inf_{y\in\Omega} U(y)\big) = 0. \quad (2)$$

This is the case for logarithmic cooling sequences of the form $\beta_n = \beta_0 \ln(n+1)$ provided β_0 is smaller than a critical value β_{c} that depends on U and θ. However, logarithmic cooling is inefficient for most practical problems; indeed, β_{c} is generally too large to reach the low temperature regime in a reasonable amount of computation time, whereas the process gets easily stuck in poor local minima for feasible values of β_0.

Designing an efficient SA algorithm means smartly choosing the communication mechanism θ and carefully selecting the cooling sequence $(\beta_n)_n$. These two levers for convergence acceleration are considered first, and our discussion about cooling sets the basis for introducing the technique of acceleration by concave energy distortion. We continue with a recently introduced generalization of SA, called stochastic continuation, in which both the energy function and the communication mechanism are allowed to vary with temperature. The paper ends with practical considerations for tuning the cooling schedule. Each topic is summarized below.

Design of the Communication Mechanism (Section 3). The design of the communication mechanism is application-dependent and hence cannot be reduced to a simple recipe, but there are general ideas that can lead to significant benefits in terms of convergence speed. We start with the standard construction scheme based on a neighborhood system that specifies the allowed moves. We then discuss three concepts that can facilitate the exploration of the state space and that can be tied together: state-space restriction, as successfully used in [28], state-space transformation, an example of which can be found in [26], and relaxation.

Finite-Time Cooling (Section 4). The issue of finite-time cooling is of primary importance, as the available computing time is always bounded in practice. We investigate the finite-time convergence results of Catoni [6], who showed that the convergence rate cannot be faster than some optimal power of $1/n$ and that exponential cooling must be preferred over logarithmic cooling. More precisely, the optimal convergence speed exponent is $1/D$, where D is the so-called *difficulty of the energy landscape* (D is a function of U and θ), and it is possible to construct a family $\{(\beta_n^N)_{1\leqslant n\leqslant N}\,;\, N \in \mathbb{N}^*\}$ of finite cooling sequences of the form $\beta_n^N = \beta_0 \exp(\zeta n)$, where $\zeta \in (0, +\infty)$ depends on N, such that

$$\ln \mathsf{P}\big(U(X_N) > \inf_{y\in\Omega} U(y)\big) \sim \ln N^{-1/D}. \quad (3)$$

These results are not well-known, and yet they constitute the most significant advance in SA theory beyond the asymptotic properties established in [16]: they provide a rigorous justification for the commonly used exponential cooling schedules.

Concave Energy Distortion (Section 5). The convergence results associated with finite-time SA ground the theoretical justification for acceleration by distortion of the energy function [28]. The technique simply consists in replacing U by $\varphi \circ U$, where the function φ is differentiable, increasing, and strictly concave. The rationale behind this is that the difficulty of the energy landscape D_φ associated with $\varphi \circ U$ is strictly smaller than the original difficulty D; therefore, the optimal convergence speed exponent is increased, thus leading to potential acceleration. We also discuss a theoretical way to compare the relative performance of different distortion functions.

Stochastic Continuation (Section 6). Stochastic continuation (SC) is a recently introduced generalization of SA which relaxes the design constraints of annealing-type algorithms by allowing the energy function and the communication mechanism to vary with temperature [24, 27, 29, 30]. The first idea is to ease the optimization process by gradually revealing its complexity, which can be obtained by replacing the energy U by a sequence of functions converging pointwise to U with increasing difficulty. The second idea is to facilitate the exploration of the state space by adapting the communication mechanism to the temperature regime. Formally, an SC algorithm is defined by a family $(U_\beta)_{\beta \in \mathbb{R}_+}$ of real-valued functions on Ω called the *continuation scheme*, a family $(\theta_\beta)_{\beta \in \mathbb{R}_+}$ of Markov matrices on Ω called the *communication scheme*, and a cooling sequence $(\beta_n)_{n \in \mathbb{N}^*}$; the description of SC is the same as that of SA, except that the energy U and the communication mechanism θ in (1) are respectively replaced by U_β and θ_β.

We give the conditions for SC to have finite-time convergence properties similar to that of SA. These conditions are surprisingly weak, and, quite interestingly, exponential cooling makes it possible for SC to have a convergence speed exponent arbitrarily close to the optimal exponent of SA. More precisely, letting D be the difficulty of the energy landscape defined by the limit energy $U = \lim_{\beta \to +\infty} U_\beta$ and by the limit communication matrix $\lim_{\beta \to +\infty} \theta_\beta$, we have that for any $\alpha \in (0, 1/D)$, there is a family $\{(\beta_n^N)_{1 \leqslant n \leqslant N}\,;\, N \in \mathbb{N}^*\}$ of finite exponential cooling sequences such that

$$\mathsf{P}\big(U(X_N) > \inf_{y \in \Omega} U(y)\big) \leqslant N^{-\alpha} \tag{4}$$

for N large enough. We end our discussion of SC with guidelines for constructing the continuation and communication schemes.

Practical Tuning of the Cooling Sequence (Section 7). The exponential cooling sequences suggested by SC theory are of the form

$$\beta_n^N = \beta_0 \exp\Big(\zeta \Big\lceil \frac{\sigma}{N} n \Big\rceil\Big), \tag{5}$$

where $\lceil \cdot \rceil$ is the ceiling function and σ is the number of constant-temperature stages. Generally, σ is fixed in advance and the horizon N is a multiple of σ that is fixed by the available computing resources. This leaves us with the problem of finding appropriate values for β_0 and ζ, or equivalently for the initial and final inverse temperatures $\beta_{\inf} := \beta_0 \exp(\zeta)$ and $\beta_{\sup} := \beta_0 \exp(\zeta\sigma)$. We discuss efficient approximate methods for estimating $\beta_{\inf}$ and $\beta_{\sup}$ according to criteria on the ratio of the number of accepted uphill moves to the number of proposed ones.

2 Simulated Annealing

We consider the problem of finding a global minimum of an arbitrary real-valued energy function U defined on a finite state space Ω. We denote the ground state energy by $U_{\inf}$, and we let $\Omega_{\inf}$ be the set of global minima of U; that is,

$$U_{\inf} = \inf_{x\in\Omega} U(x) \qquad \text{and} \qquad \Omega_{\inf} = \big\{x \in \Omega \,\big|\, U(x) = U_{\inf}\big\}. \tag{6}$$

Given two integers a and b such that $a \leqslant b$, we denote by $[\![a,b]\!]$ the set of integers between a and b, including a and b. This notation will be used throughout the paper.

2.1 Fundamentals

Simulated annealing (SA) operates on an *energy landscape* (Ω, U, θ) defined by a symmetric and irreducible Markov matrix θ on Ω, called the *communication matrix*, which specifies how to generate a candidate solution from the current solution. More precisely, we assume that $\theta : \Omega^2 \to [0,1]$ has the following properties.

1. θ is a Markov matrix: $\sum_{z\in\Omega} \theta(x,z) = 1$ for all $x \in \Omega$.
2. θ is symmetric: $\theta(x,y) = \theta(y,x)$ for all $(x,y) \in \Omega^2$.
3. θ is irreducible: for any $(x,y) \in \Omega^2$, there is a θ-*admissible path* from x to y, that is, a path $(x_i)_{i=1}^m$ such that $x_1 = x$, $x_m = y$, and $\theta(x_i, x_{i+1}) > 0$ for all $i \in [\![1, m-1]\!]$.

In simple terms, the probability to propose a move from x to y is the same as that to propose a move from y to x, and any state can be reached from any other state in a finite number of moves. (Standard and advanced construction schemes for θ are described in Section 3.)

An SA process on an energy landscape (Ω, U, θ) is defined by a family $(P_\beta)_{\beta\in\mathbb{R}_+}$ of Markov matrices on Ω of the form

$$P_\beta(x,y) = \begin{cases} \theta(x,y) A_\beta(x,y) & \text{if } y \neq x, \\ 1 - \sum_{z\in\Omega\setminus\{x\}} P_\beta(x,z) & \text{if } y = x, \end{cases} \tag{7}$$

where the so-called *acceptance probability function* $A_\beta : \Omega^2 \to [0,1]$ is defined by

$$A_\beta(x,y) = \exp\big(-\beta(U(y) - U(x))^+\big) \tag{8}$$

with $t^+ := \sup\{t, 0\}$. The parameter β plays the role of an inverse temperature, and $A_\beta(x,y)$ is the probability to accept the move from the current solution x to the candidate solution y at temperature β^{-1}. Other acceptance probability functions are possible (we then speak of an hill-climbing process [17]), but it is shown in [32] that (8) is the unique form such that (i) $A_\beta(x,y) = 1$ if $U(y) \leqslant U(x)$, (ii) A_β depends uniformly on the energy difference between the current and candidate solutions, and (iii) the Markov chain $(X_n)_{n\in\mathbb{N}}$ with transitions $\mathsf{P}(X_n = y \,|\, X_{n-1} = x) = P_\beta(x,y)$ is reversible.

We call a positive real sequence $(\beta_n)_{n\in\mathbb{N}^*}$ a *cooling sequence* if it is non-decreasing and if $\lim_{n\to+\infty}\beta_n = +\infty$. Given such a sequence, an *SA algorithm* on (Ω, U, θ) is a discrete-time, non-homogeneous Markov chain $(X_n)_{n\in\mathbb{N}}$ with transitions $\mathsf{P}(X_n = y \,|\, X_{n-1} = x) = P_{\beta_n}(x,y)$. We use the notation $\mathrm{SA}(\Omega, U, \theta, (\beta_n))$ for short. In practice, a finite-time realization $(x_n)_{n\in[\![0,N]\!]}$ of an annealing chain $\mathrm{SA}(\Omega, U, \theta, (\beta_n))$ is generated as follows:

```
pick an initial state $x_0 \in \Omega$;
**for** $n = 1$ **to** $N$ **do**
    draw a state $y$ from the probability distribution $\theta(x_{n-1}, \cdot)$ on $\Omega$;
    set $x_n \longleftarrow x_{n-1}$;
    set $\delta \longleftarrow U(y) - U(x_{n-1})$;
    **if** $\delta \leqslant 0$ **then** set $x_n \longleftarrow y$;
    **else** set $x_n \longleftarrow y$ with probability $\exp(-\beta_n\delta)$;
    **end(if)**
**end(for)**
```

The Markov matrix P_β inherits the irreducibility of θ for any β. Therefore, since Ω is finite, P_β has a unique and positive invariant measure which we denote by μ_β. Moreover, from the symmetry of θ, we have

$$\exp(-\beta U(x))P_\beta(x,y) = \exp(-\beta U(y))P_\beta(y,x) \tag{9}$$

for all $(x,y) \in \Omega^2$, that is, P_β is reversible with respect to a distribution proportional to $\exp(-\beta U(x))$, and thus

$$\mu_\beta(x) = \frac{\exp(-\beta U(x))}{\sum_{z\in\Omega}\exp(-\beta U(z))} \tag{10}$$

for all $x \in \Omega$. In other words, the steady-state distribution of P_β is the Gibbs distribution with energy U at temperature β^{-1}. When β increases to infinity, this distribution concentrates around the ground states and tends to the uniform distribution on $\Omega_{\inf}$, that is,

$$\lim_{\beta\to+\infty}\mu_\beta(x) = \begin{cases} 1/|\Omega_{\inf}| & \text{if } x \in \Omega_{\inf}, \\ 0 & \text{if } x \notin \Omega_{\inf}. \end{cases} \tag{11}$$

This observation leads to the key idea of annealing: if the cooling sequence $(\beta_n)_n$ increases sufficiently slowly, then we can expect that the law of X_n stays close enough to μ_{β_n} so that

$$\lim_{n\to+\infty}\inf_{x\in\Omega}\mathsf{P}\big(X_n \in \Omega_{\inf} \,\big|\, X_0 = x\big) = 1. \tag{12}$$

However, it is natural to question the need for cooling. Indeed, we can think of searching for the global minima by Metropolis sampling, which consists in simulating an homogeneous Markov chain with transitions matrix P_β for a fixed β and keeping the lowest energy state found during the simulation. Metropolis sampling has interesting finite-time convergence properties [7, 12, 23, 22], and some experimental results show that it can perform comparably to SA if the temperature is chosen correctly [10, 13]. Unfortunately, there is no general approach to choosing a fixed temperature value appropriate to a given optimization problem. The difficulty is the following. On the one hand, if we want to be reasonably sure of finding a good solution, we have to choose β large enough so that μ_β is sharply peaked around the ground states. On the other hand, the larger β, the less mobile the Metropolis chain, and hence the more likely it is to get stuck in poor local minima. From this perspective, SA can be viewed as an acceleration technique for Metropolis sampling.

2.2 Asymptotic Convergence

The most well-known asymptotic convergence result for SA is due to Hajek [16], who showed that (12) holds if and only if

$$\sum_{n=1}^{+\infty} \exp(-\beta_n H_\mathrm{c}) \;=\; +\infty, \tag{13}$$

where H_c is the maximum energy barrier separating a non-optimal state from a ground state. The constant H_c is called the *critical depth of the energy landscape*. Formally,

$$H_\mathrm{c} \;=\; \sup_{x\in\Omega\setminus\Omega_\mathrm{inf}} H(x), \tag{14}$$

where $H(x)$ — the *depth of* x — is defined as follows:

$$H(x) \;=\; \inf_{y\in\Omega_\mathrm{inf}} h(x,y) - U(x) \tag{15}$$

$$\text{with} \qquad h(x,y) \;=\; \inf_{(x_i)_{i=1}^m \in \Pi_\theta(x,y)} \; \sup_{i\in[\![1,m]\!]} U(x_i), \tag{16}$$

where $\Pi_\theta(x,y)$ denotes the set of θ-admissible paths from x to y.

Hajek's result readily implies that logarithmic cooling sequences of the form $\beta_n = \beta_0 \ln(n+1)$ are asymptotically optimal if $0 < \beta_0 \leqslant 1/H_\mathrm{c}$. A notable refinements was given by Chiang and Chow [9] who provided a necessary and sufficient condition for the limit distribution of the annealing chain to give a strictly positive mass to any global minimum: assuming that $|\Omega_\mathrm{inf}| \geqslant 2$, we have

$$\Big\{ y\in\Omega \;\Big|\; \lim_{n\to+\infty} \inf_{x\in\Omega} \mathsf{P}\big(X_n = y \,\big|\, X_0 = x\big) \;>\; 0 \Big\} \;=\; \Omega_\mathrm{inf} \tag{17}$$

if and only if

$$\sum_{n=1}^{+\infty} \exp(-\beta_n \sup\{H_{\mathrm{c}}, H_{\mathrm{inf}}\}) \;=\; +\infty, \tag{18}$$

where H_{inf} is the maximum energy barrier separating two ground states, that is,

$$H_{\mathrm{inf}} \;=\; \sup_{(x,y)\in\Omega_{\mathrm{inf}}^2} h(x,y) - U_{\mathrm{inf}}. \tag{19}$$

A necessary and sufficient condition for strong ergodicity can be found in [5]. This condition is similar to (18) but with a critical constant greater than or equal to $\sup\{H_{\mathrm{c}}, H_{\mathrm{inf}}\}$, and it ensures that for any $x \in \Omega$,

$$\lim_{n\to+\infty} \mathsf{P}\big(X_n = y \,\big|\, X_0 = x\big) \;=\; \begin{cases} 1/|\Omega_{\mathrm{inf}}| & \text{if } y \in \Omega_{\mathrm{inf}}, \\ 0 & \text{if } y \notin \Omega_{\mathrm{inf}}. \end{cases} \tag{20}$$

However, these asymptotic results impose logarithmic cooling, which yields extremely slow convergence, while successful applications of SA generally use exponential cooling. Furthermore, convergence guarantees such has (12), (17) or (20) are of limited interest if the horizon is finite, as is always the case in practice. The finite-time convergence properties of SA along with the justification of exponential cooling are discussed in Section 4.

3 Design of the Communication Mechanism

The communication mechanism is usually defined via a *neighborhood system* $\mathscr{G}$ on Ω, that is, a collection $\mathscr{G} = \{\mathscr{G}(x)\,;\, x \in \Omega\}$ of subsets of Ω such that (i) $x \notin \mathscr{G}(x)$ for all $x \in \Omega$, and (ii) $y \in \mathscr{G}(x) \iff x \in \mathscr{G}(y)$ for all $(x,y) \in \Omega^2$. We let $\Delta(\mathscr{G}) = \big\{\{x,y\} \subset \Omega \,\big|\, y \in \mathscr{G}(x)\big\}$ be the set of neighboring state pairs in $\mathscr{G}$ and $\big(\Omega, \Delta(\mathscr{G})\big)$ be the adjacency graph with vertex set Ω and edge set $\Delta(\mathscr{G})$. The most simple mechanisms have the following form:

$$\theta(x,y) \;=\; \begin{cases} c & \text{if } y \in \mathscr{G}(x), \\ 1 - c\,|\mathscr{G}(x)| & \text{if } y = x, \\ 0 & \text{otherwise,} \end{cases} \tag{21}$$

with $0 < c \leqslant 1/\big(\sup_{x\in\Omega} |\mathscr{G}(x)|\big)$. A standard example is that of a single-component updating communication mechanism on a cartesian product space $\Omega = \Upsilon^d$. In this case, a candidate solution $y = (y_1, \ldots, y_d)$ is generated from $x = (x_1, \ldots, x_d)$ by picking a component index $i \in [\![1,d]\!]$ and a component value $t \in \Upsilon$ uniformly at random and setting $y_i = t$ and $y_j = x_j$ for all $j \neq i$. The associated communication matrix writes

$$\theta(x,y) = \begin{cases} 1/(d|\Upsilon|) & \text{if } \exists! i \in [\![1,d]\!],\ y_i \neq x_i, \\ 1/|\Upsilon| & \text{if } y = x, \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

More sophisticated mechanisms are constructed by weighting the allowed moves using a function $\gamma : \Delta(\mathscr{G}) \to (0,+\infty)$; they are of the form

$$\theta(x,y) = \begin{cases} c\,\gamma(\{x,y\}) & \text{if } y \in \mathscr{G}(x), \\ 1 - c \sum_{z \in \mathscr{G}(x)} \gamma(\{x,z\}) & \text{if } y = x, \\ 0 & \text{otherwise,} \end{cases} \tag{23}$$

with $0 < c \leqslant 1/\big(\sup_{x\in\Omega} \sum_{z\in\mathscr{G}(x)} \gamma(\{x,z\})\big)$. A communication matrix of this type is clearly symmetric, and it is irreducible if and only if $\big(\Omega, \Delta(\mathscr{G})\big)$ is connected. Conversely, any symmetric and irreducible Markov matrix on Ω is of the form of (23) with $\big(\Omega, \Delta(\mathscr{G})\big)$ connected: it suffices to set $\mathscr{G}(x) = \{y \in \Omega \,|\, y \neq x \text{ and } \theta(x,y) > 0\}$ for all $x \in \Omega$, $\gamma(\{x,y\}) = \theta(x,y)$ for all $\{x,y\} \in \Delta(\mathscr{G})$, and $c = 1$. The choice of the neighborhood system $\mathscr{G}$ and of the weighting function γ depends on the structure of the optimization problem under consideration, but there are general ideas that can significantly improve the convergence speed of SA. These concepts — namely, restriction of the state-space, transformation of the state-space, and relaxation — are described below; they can be used independently or together.

3.1 Restriction of the State Space

The difficulty of minimizing a particular energy function U depends on the set Ω on which it is defined. In particular, if the solutions of interest belong to a relatively small fraction $\widetilde{\Omega}$ of Ω that can be easily identified, then it makes sense to try to minimize the energy function over $\widetilde{\Omega}$ instead of Ω. Such a *restriction* of the minimization domain is a valuable option if $|\widetilde{\Omega}| \ll |\Omega|$ and if $\widetilde{\Omega}$ is both easy to explore and rich enough to contain most acceptable solutions.

An important case is when Ω is a cartesian product space $\prod_{i=1}^{d} \Omega_i$ and $\widetilde{\Omega}$ consists of the states $x = (x_1, \ldots, x_d)$ such that each component x_i belongs to a subset of Ω_i defined as a function of the other components; that is,

$$\widetilde{\Omega} = \big\{x = (x_1, \ldots, x_d) \,\big|\, \forall i \in [\![1,d]\!],\ x_i \in F_i(x_{\setminus\{i\}})\big\}, \tag{24}$$

where each F_i is a function from $\Omega_{\setminus\{i\}} := \prod_{j=1, j\neq i}^{d} \Omega_j$ to the power set of Ω_i, and where $x_{\setminus\{i\}}$ denotes the $(d-1)$-tuple obtained by removing the ith component x_i from x. The design of a single-component updating communication mechanism $\widetilde{\theta}$ operating on $\widetilde{\Omega}$ is conceptually simple. For any $x \in \widetilde{\Omega}$, we let $x_{\setminus\{i\}}(t) \in \Omega$ be the state obtained by replacing the ith component of x by t, and we denote the section of $\widetilde{\Omega}$ at $x_{\setminus\{i\}}$ by $\omega_i(x)$; that is,

$$\forall j \in [\![1,d]\!], \quad (x_{\setminus\{i\}}(t))_j = \begin{cases} t & \text{if } j = i, \\ x_j & \text{otherwise,} \end{cases} \tag{25}$$

$$\text{and} \qquad \omega_i(x) = \big\{ t \in \Omega_i \,\big|\, x_{\setminus\{i\}}(t) \in \widetilde{\Omega} \big\}. \tag{26}$$

Then, a candidate solution y can be generated from x by picking a component index $i \in [\![1,d]\!]$ and a component value $t \in \omega_i(x)$ uniformly at random and setting $y = x_{\setminus\{i\}}(t)$. The corresponding formal description is the following:

$$\widetilde{\theta}(x,y) = \begin{cases} \dfrac{1}{d|\omega_i(x)|} & \text{if } y_{\setminus\{i\}} = x_{\setminus\{i\}} \text{ and } y_i \in \omega_i(x) \setminus \{x_i\}, \\ \dfrac{1}{d} \displaystyle\sum_{i \in [\![1,d]\!] \,:\, x_i \in \omega_i(x)} \dfrac{1}{|\omega_i(x)|} & \text{if } y = x, \\ 0 & \text{otherwise,} \end{cases} \tag{27}$$

which is of the form of (23) with

$$\mathscr{G}(x) = \bigcup_{i=1}^{d} \big\{ x_{\setminus\{i\}}(t) \,;\, t \in \omega_i(x) \setminus \{x_i\} \big\}, \tag{28}$$

$\gamma(\{x,y\}) = 1/|\omega_i(x)|$, and $c = 1/d$. The efficiency of this communication mechanism depends on how difficult it is to evaluate the ω_i's, which in turn depends on the choice of the functions F_i that define $\widetilde{\Omega}$. This choice cannot be arbitrary: it must guarantee that the adjacency graph with vertex set $\widetilde{\Omega}$ and edge set $\big\{\{x,y\} \subset \widetilde{\Omega} \,\big|\, y \in \mathscr{G}(x)\big\}$ defined by (28) is connected. A clear-cut example can be found in [28], where Ω is a digital image space and $\widetilde{\Omega}$ is a so-called *locally bounded image space* which consists of the images in which each pixel value is bounded by the values of neighboring pixels up to an additive constant.

3.2 Transformation of the State Space

The design of an efficient communication mechanism can be facilitated by transforming the domain in which the state space Ω lies. The idea is to get around the difficulty of constructing a sophisticated communication mechanism in the original minimization domain by operating on a transformed domain that can be effectively explored using a simple communication mechanism.

By way of illustration, consider the case when Ω is a cartesian product space indexed by the sites of a spatial lattice, as in image processing. If the lattice has a large number of sites, it is a common situation that the energy bonds between the state components are loose and hence that SA with single-component updating experiences difficulties (see, for instance, Jennison's discussion in [1]). This is especially true when the low energy regions of the state space correspond to smooth configurations, as moving between such regions by changing only one component at a time requires many iterations. The obvious answer to this problem is to generate candidate solutions by changing several components simultaneously, but direct design of a multiple-component updating mechanism can be very cumbersome. An effective way to do that is to operate in a multiresolution transform domain, because single-component updating at a coarse resolution level corresponds to multiple-component updating in the finest scale (that is, in the original domain) and hence improves the mobility of the annealing chain. A comprehensive example is given in [26], where multi-component updating is achieved by performing single-component moves in a wavelet transform domain.

Formally, the state-space transformation approach uses a bijective map between the original domain, say Λ, which contains Ω, and the transformed domain which we denote by $\widetilde{\Lambda}$. Let $\pi : \Lambda \to \widetilde{\Lambda}$ be such a map, and assume for the sake of generality that Λ (and hence $\widetilde{\Lambda}$) is uncountable. Then, since SA operates on finite state spaces, $\widetilde{\Lambda}$ must be restricted to a finite set $\widetilde{\Omega}$ on which an efficient communication mechanism can be easily constructed. If $\pi^{-1}(\widetilde{\Omega}) \subseteq \Omega$, the original problem of minimizing U over Ω is replaced by that of minimizing $U \circ \pi^{-1}$ over $\widetilde{\Omega}$. However, it can be difficult to ensure that $\pi^{-1}(\widetilde{\Omega}) \subseteq \Omega$, and thus, strictly speaking, the new optimization problem is that of minimizing $U|^{\Lambda} \circ \pi^{-1}$ over $\widetilde{\Omega}$, where $U|^{\Lambda}$ is an extension of U to Λ. This brings us to the concept of relaxation.

3.3 Relaxation

Extending the set Ω over which the energy U is to be minimized is called a *relaxation* of the original problem; it is adapted to situations where (i) the structure of Ω complicates the optimization problem unnecessarily, and (ii) there is a larger set $\Lambda \supset \Omega$ that contains interesting approximate solutions that can be found more easily than the global minima of U on Ω.

Given an extension $U|^{\Lambda}$ of the original energy to Λ, we denote the ground state energy $\inf_{x\in\Lambda}(U|^{\Lambda})(x)$ by $(U|^{\Lambda})_{\text{inf}}$ and we let Λ_{inf} be the set of global minima of $U|^{\Lambda}$. Ideally, $(U|^{\Lambda})_{\text{inf}} = U_{\text{inf}}$ and there exists a surjective map $\kappa : \Lambda \to \Omega$ such that $\kappa(\Lambda_{\text{inf}}) = \Omega_{\text{inf}}$, so that solving the relaxed problem solves the original problem. Otherwise, the computed solution only provides a lower bound on U_{inf}. In practice, however, one is usually satisfied with solutions whose energy level is close to the ground state energy rather than with global minima only; that is, the set of solutions is extended from Ω_{inf} to a sublevel set

$$\Omega_{\varepsilon} = \left\{ x \in \Omega \,\middle|\, U(x) \leqslant U_{\text{inf}} + \varepsilon \right\}, \tag{29}$$

where $\varepsilon > 0$ is a given tolerance level. In this case, a basic requirement is that κ maps the acceptable solutions to the relaxation to acceptable solutions to the original problem, that is,

$$\forall \varepsilon > 0, \quad \exists \alpha > 0, \quad \kappa(\Lambda_{\alpha}) \subseteq \Omega_{\varepsilon}, \tag{30}$$

where $\Lambda_{\alpha} = \left\{ x \in \Lambda \mid (U|^{\Lambda})(x) \leqslant (U|^{\Lambda})_{\text{inf}} + \alpha \right\}$.

Relaxation is the opposite concept to restriction, but both can be used together when the minimization is performed in a transformed domain, as summarized by the following diagram:

$$\begin{tikzcd}
\Omega & \widetilde{\Omega} \arrow[d, "\textbf{Restriction}", "\widetilde{\iota}\ \text{(inclusion map)}"'] \\
\Lambda \arrow[u, "\textbf{Relaxation}", "\kappa\ \text{(surjection)}"'] \arrow[r, "\pi\ \text{(bijection)}", "\textbf{Transformation}"'] & \widetilde{\Lambda}
\end{tikzcd} \tag{31}$$

Let $\widetilde{\Omega}_{\text{inf}}$ be the set of solutions of the transformed optimization problem, that is, the set of global minima of the transformed energy $U|^{\Lambda} \circ \pi^{-1}$ over the restricted set $\widetilde{\Omega}$. Finding a state in $\widetilde{\Omega}_{\text{inf}}$ solves the original problem of minimizing U over Ω if and only if the set $\Omega'_{\text{inf}} := \kappa(\pi^{-1}(\widetilde{\Omega}_{\text{inf}}))$ is a subset of Ω_{inf}. Otherwise, the original solution set Ω_{inf} is implicitly replaced by the approximate solution set Ω'_{inf}, which makes sense if the set of acceptable solutions is of the form of (29) and if κ satisfies (30).

4 Finite-Time Cooling

Given an energy landscape (Ω, U, θ) and a finite cooling sequence $(\beta_n^N)_{n \in [\![1,N]\!]}$, we define the convergence measure $\mathsf{M}(N)$ of the finite-time annealing algorithm $(X_n^N)_{n \in [\![0,N]\!]} = \mathsf{SA}(\Omega, U, \theta, (\beta_n^N))$ by

$$\mathsf{M}(N) = \sup_{x \in \Omega} \mathsf{P}\big(X_N^N \notin \Omega_{\text{inf}} \,\big|\, X_0^N = x\big). \tag{32}$$

It is shown in [6] that as the horizon N increases, $\mathsf{M}(N)$ cannot decrease faster than some optimal exponent of N^{-1}. More precisely, let $\mathfrak{B}(N)$ be the set of finite cooling sequences of length N, that is, $\mathfrak{B}(N) = \{(\beta_n^N) \mid 0 \leqslant \beta_1^N \leqslant \cdots \leqslant \beta_N^N\}$. We have

$$\lim_{N \to +\infty} \sup_{(\beta_n^N) \in \mathfrak{B}(N)} -\frac{\ln \mathsf{M}(N)}{\ln N} \leqslant \frac{1}{D}, \tag{33}$$

where D denotes the *difficulty of the energy landscape*, which is the maximum ratio of the depth to the energy level above the ground state energy:

$$D = \sup_{x \in \Omega \setminus \Omega_{\text{inf}}} \frac{H(x)}{U(x) - U_{\text{inf}}}. \tag{34}$$

Furthermore, the upper bound $1/D$ in (33) is sharp, as there are some families $\{(\beta_n^N)_{n \in [\![1,N]\!]}\,;\, N \in \mathbb{N}^*\}$ of finite exponential cooling sequences such that

$$\lim_{N \to +\infty} -\frac{\ln \mathsf{M}(N)}{\ln N} = \frac{1}{D}, \tag{35}$$

which implies in particular that for any $\alpha \in (0,1/D)$, $\mathsf{M}(N) \leqslant N^{-\alpha}$ for N large enough. These families are of the form

$$\beta_n^N = \beta_0 \exp(nf(N)) \qquad \text{with} \qquad f(N) \sim N^{-1} \ln N, \tag{36}$$

where $\beta_0 \in (0,+\infty)$ is independent of N. This rigorous justification for exponential cooling is a direct consequence of Theorem 8.1 in [6] (see [28]), where it is also established that there exists piecewise logarithmic sequences such that $\mathsf{M}(N) \leqslant CN^{-1/D}$ for some positive constant C (however, these sequences depend strongly on the hierarchical structure of the energy landscape and their identification is intractable for problems of practical size). On the experimental side, the optimal cooling sequence attached to a particular optimization problem may be neither logarithmic nor exponential [10], but exponential cooling is particularly attractive because, contrary to other cooling strategies, it is uniformly robust with respect to the energy landscape.

It can be checked that the supremum in the definition (34) of the difficulty of the energy landscape can be taken over the set of non-global minima of (Ω,U,θ), that is, over

$$\Omega_{\mathrm{loc}}^{\dagger} = \Omega_{\mathrm{loc}} \setminus \Omega_{\mathrm{inf}}, \tag{37}$$

where Ω_{loc} denotes the set of local minima of (Ω,U,θ):

$$\Omega_{\mathrm{loc}} = \big\{x \in \Omega \bigm| \forall y \in \Omega,\ \theta(x,y) > 0 \Longrightarrow U(x) \leqslant U(y)\big\}. \tag{38}$$

Therefore, the above finite-time convergence properties are consistent with the intuitive understanding of annealing, that is, that SA performs poorly if the energy landscape has low-energy non-global minima and if these minima are separated from the ground states by high energy barriers. It should be stressed that this understanding differs from that stemming from the asymptotic convergence results of Hajek exposed in Section 2.2. Indeed, the supremum in the definition (14) of the critical height H_{c} can also be taken on $\Omega_{\mathrm{loc}}^{\dagger}$, and thus the asymptotic performance of SA is dictated by the maximum energy barrier separating a non-global minimum from a global one regardless of their relative energies. By way of illustration, Fig. 1 shows three simple energy landscapes with increasing difficulty. In each case, $\Omega = \{x_i\,;\, i \in [\![1,12]\!]\}$, $U(\Omega) \subset \mathbb{N}$, and $\theta(x,y) > 0$ if and only if $(x,y) = (x_i,x_{i+1})$ or (x_i,x_{i-1}). The quantities η_1 and η_2 are defined by

$$\eta_1 = H(x^*) \qquad \text{and} \qquad \eta_2 = U(x^*) - U_{\mathrm{inf}} \tag{39}$$

$$\text{with} \qquad x^* \in \operatorname*{arg\,sup}_{x \in \Omega \setminus \Omega_{\mathrm{inf}}} \frac{H(x)}{U(x) - U_{\mathrm{inf}}}, \tag{40}$$

and thus $D = \eta_1/\eta_2$. As exemplified by Figs. 1(a) and 1(c), the non-global minimum with maximum depth does not necessarily coincide with the argument of the supremum in the definition of the difficulty. The reason is that the ordering of the non-global minima in terms of the depth H is generally not the same as that

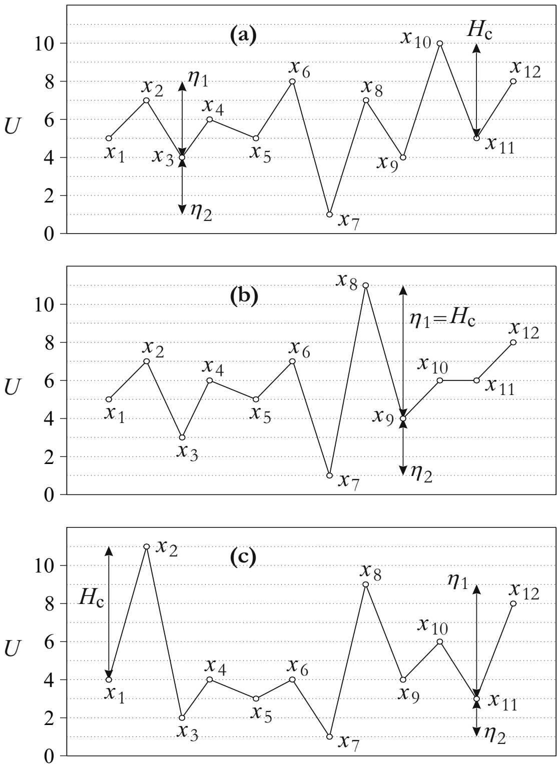

Fig. 1 Energy landscapes with increasing difficulty $D = \eta_1/\eta_2$: (a) $D = \frac{4}{3}$; (b) $D = \frac{7}{3}$; (c) $D = 3$.

defined by $H\big/(U - U_{\text{inf}})$, which means in particular that the notion of a local basin of attraction differs between asymptotic and finite-time convergence theories.

The finite-time convergence theory also sheds new light on the benefits of SA over Metropolis sampling. From [7], the optimal convergence speed exponent of the Metropolis algorithm is $1/D_{\text{M}}$ with

$$D_{\text{M}} = \frac{H_{\text{c}}}{\inf\limits_{x \in \Omega \setminus \Omega_{\text{inf}}} U(x) - U_{\text{inf}}}. \tag{41}$$

We have $D < D_{\text{M}}$ if and only if one of the following two conditions holds:

1. there exists $x \in \Omega$ such that $U_{\text{inf}} < U(x) < \inf_{y \in \Omega^{\dagger}_{\text{loc}}} U(y)$;
2. for any $x \in \Omega^{\dagger}_{\text{loc}}$, $H(x) = \sup_{y \in \Omega^{\dagger}_{\text{loc}}} H(y) \implies U(x) > \inf_{y \in \Omega^{\dagger}_{\text{loc}}} U(y)$.

In other words, SA is potentially faster than Metropolis sampling if there is a state $x \notin \Omega_{\text{inf}}$ with smaller energy than any non-global minimum or if the set of non-global minima with maximum depth is disjoint from the set of non-global minima with minimum energy. For example, going back to Fig. 1, we have $D < D_{\mathrm{M}}$ in all three cases: (a) $D_{\mathrm{M}} = \frac{5}{3}$, (b) $D_{\mathrm{M}} = \frac{7}{2}$, and (c) $D_{\mathrm{M}} = 7$.

5 Concave Energy Distortion

We know from finite-time SA theory (see Section 4) that there are some families of exponential cooling sequences such that $\mathsf{M}(N)$ is asymptotically equivalent to $N^{-1/D}$ in the logarithmic scale, where $1/D$—the inverse of the difficulty of the energy landscape—is the optimal convergence speed exponent. Therefore, the more difficult the energy landscape (as measured by D), the lower the convergence rate, and we can ask ourselves whether there are convenient ways to reduce the difficulty without changing the set of solutions of the underlying minimization problem. The concave distortion idea proposed by Azencott [3, 2] makes this possible.

Let φ be a strictly increasing function defined on an interval covering the range of U. Then the set of global minima of $\varphi \circ U$ is the same as the set of global minima of U, and thus the minimization of U can be performed equally well by replacing U with $\varphi \circ U$. The nice thing is that if, in addition, φ is strictly concave, then the difficulty $D(\Omega, \varphi \circ U, \theta)$ of the distorted energy landscape is smaller than the difficulty $D(\Omega, U, \theta)$ of the original energy landscape, which means that annealing algorithms of type $\mathrm{SA}(\Omega, \varphi \circ U, \theta, (\beta_n))$ are expected to converge faster than those of type $\mathrm{SA}(\Omega, U, \theta, (\beta_n))$. This is made precise by the following theorem whose proof is given in [28].

Theorem 1. *Let I be an open interval covering the range of U. For any increasing, strictly concave, differentiable function $\varphi : I \to \mathbb{R}$, the set of global minima of $\varphi \circ U$ is the same as the set of global minima of U and $D(\Omega, \varphi \circ U, \theta) < D(\Omega, U, \theta)$.*

Roughly speaking, the idea is that an increasing concave transform of the energy function exaggerates the depth of global minima. As an example, consider the energy landscape shown in Fig. 1(c), which has a difficulty of 3. Using a logarithmic transform, we obtain the energy landscape $(\Omega, \ln U, \theta)$ displayed in Fig. 2(a). The global minimum x_7 appears deeper compared to the local minima x_1, x_3, x_5, x_9 and x_{11}, and hence the chance of getting stuck in a non-optimal state is reduced. Quantitatively, the distorted energy landscape has a difficulty of 1, and thus the maximum acceleration is of the order of $N^{-1/3}/N^{-1} = N^{2/3}$. This effect is even more pronounced when using $\varphi(u) = -\exp(-u)$, as shown in Fig. 2(b): the difficulty of the distorted energy landscape $(\Omega, -\exp(-U), \theta)$ is close to $\frac{1}{2}$, and the maximum acceleration is of the order of about $N^{5/3}$.

Some example functions for energy distortion are

$$\varphi_1^{\tau,a}(u) = (u-a)^{1/\tau}, \qquad \tau \in (1,+\infty), \tag{42}$$

$$\varphi_2^{\tau,a,b}(u) = \ln\big((b-a)^{\tau} - (b-u)^{\tau}\big), \qquad \tau \in [1,+\infty), \tag{43}$$

and

$$\varphi_3^{\tau,a}(u) = -\exp\big(-\tau(u-a)\big), \qquad \tau \in (0,+\infty), \tag{44}$$

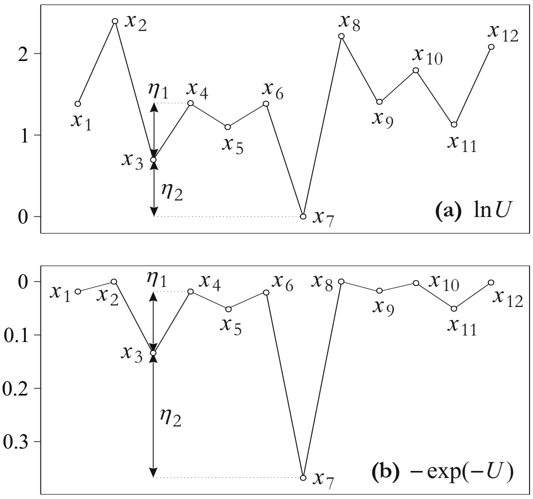

Fig. 2 Increasing concave transforms of the energy landscape shown in Fig. 1(c): (a) $D(\Omega, \ln U, \theta) = 1$; (b) $D(\Omega, -\exp(-U), \theta) \approx 0.503$.

where $a \in (-\infty, U_{\text{inf}})$ and $b \in (U_{\text{sup}}, +\infty)$ with $U_{\text{sup}} := \sup_{x \in \Omega} U(x)$. The problem of choosing suitable values for the parameters τ, a and b, along with the fact that many other families of concave transforms are conceivable, raises the question of whether a theoretical mean of comparison can be found. We have the following result which encourages the use of functions with large concavity-to-increase ratio (see [28] for proof).

Theorem 2. *Let I be an open interval covering the range of U, and let φ and ψ be twice-differentiable increasing functions from I to $\mathbb{R}$. If*

$$\left(-\frac{\varphi''}{\varphi'}\right)(u) < \left(-\frac{\psi''}{\psi'}\right)(u) \qquad \textit{for all } u \in I, \tag{45}$$

then $D(\Omega, \psi \circ U, \theta) < D(\Omega, \varphi \circ U, \theta)$.

Putting it simply, ψ is a "better" concave transform than φ if condition (45) holds, which we denote by $\varphi \prec \psi$. For instance, considering examples (42) and (43), we have, for any $a < U_{\text{inf}}$ and any $b > U_{\text{sup}}$,

$$\forall (\tau, \tau') \in [1, +\infty)^2, \quad \begin{cases} \tau < \tau' \implies \varphi_1^{\tau,a} \prec \varphi_1^{\tau',a}, \\ \varphi_1^{\tau,a} \prec \varphi_2^{\tau',a,b}, \\ \tau < \tau' \implies \varphi_2^{\tau,a,b} \prec \varphi_2^{\tau',a,b}. \end{cases} \tag{46}$$

Moreover, the closer a and b are to U_{inf} and U_{sup}, the larger the concavity-to-increase ratio, and thus the higher the potential acceleration. Note, however, that trying to find the best possible distortion function in terms of the strict order $\prec$ may not be fruitful: for instance, although the functions $\varphi_3^{\tau,a}$ defined in (44) have the remarkable property that $-(\varphi_3^{\tau,a})''/(\varphi_3^{\tau,a})' = \tau$, and hence virtually unbounded acceleration capability, they are practically unfeasible even for small values of τ. Experiments demonstrating the benefits of concave energy distortion can be found in [28] and [25], where we focus on typical optimization problems in image restoration and in image reconstruction from line-integral projections.

6 Stochastic Continuation

The relationship between the convergence rate of SA and the difficulty of the energy landscape suggests a possible acceleration by making the energy temperature-dependent. The idea is to guide the hierarchical search performed by SA — and thus to reduce the risk of getting stuck in undesirable basins of attraction — by replacing the energy function U with a sequence $(U_n)_n$ of functions converging pointwise to U and such that the difficulty of (Ω, U_n, θ) increases with n. In a similar vein, since static communication is generally efficient over only a small range of temperatures, another potential improvement is to adapt the communication mechanism to the temperature regime. This leads to an important class of generalized SA algorithms in which the temperature controls not only the acceptance rate of uphill moves, but also the energy function and the communication matrix. We call it stochastic continuation (SC) by analogy with deterministic continuation methods, in which the minima are tracked by computing successive approximate solutions from a parameterized energy that tends to the objective function as the iterations increase (see, for instance, [4, 20, 21]).

6.1 Definition and Basic Idea

In a nutshell, SC is a variant of SA in which both the energy function and the communication mechanism can vary with temperature. More precisely, we define an SC process with target energy landscape (Ω, U, θ) to be a family $(Q_\beta)_{\beta\in\mathbb{R}_+}$ of Markov matrices on Ω of the form

$$Q_\beta(x,y) = \begin{cases} \theta_\beta(x,y)\exp\bigl(-\beta(U_\beta(y) - U_\beta(x))^+\bigr) & \text{if } y \neq x, \\ 1 - \sum_{z\in\Omega\setminus\{x\}} Q_\beta(x,z) & \text{if } y = x, \end{cases} \tag{47}$$

$$\text{with} \qquad \lim_{\beta\to+\infty} U_\beta(x) = U(x) \qquad \text{and} \qquad \lim_{\beta\to+\infty} \theta_\beta(x,y) = \theta(x,y).$$

Given such a family together with a cooling sequence $(\beta_n)_{n\in\mathbb{N}^*}$, we call a Markov chain $(X_n)_{n\in\mathbb{N}}$ on Ω with transitions $\mathsf{P}(X_n = y \,|\, X_{n-1} = x) = Q_{\beta_n}(x,y)$ an *SC algorithm*, and we denote it by $\mathrm{SC}(\Omega, (U_\beta), (\theta_\beta), (\beta_n))$. The family of functions $(U_\beta : \Omega \to \mathbb{R})_\beta$ is called the *continuation scheme*, and the family of Markov matrices $(\theta_\beta : \Omega^2 \to [0,1])_\beta$ is called the *communication scheme*.

The limit communication matrix θ is assumed to be irreducible, as otherwise the target energy landscape cannot be freely explored and there is no guarantee to reach a ground state of the target energy U. The basic idea of SC is similar to that of SA and is quite easy to explain if θ_β is symmetric for all β (this assumption is relaxed in the next section). Indeed, in this case, the invariant measure ν_β of Q_β is the Gibbs distribution with energy U_β at temperature β^{-1}, that is, $\nu_\beta(x) \propto \exp(-\beta U_\beta(x))$, and this distribution concentrates on the set $\Omega_{\inf}$ of global minima of U as $\beta \to +\infty$ [29]. Consequently, similarly to SA, if the cooling sequence does not increase too fast, the law of X_n should stay close enough to ν_{β_n} to expect convergence to an optimum.

6.2 Finite-Time Convergence

SC is an extension of SA with temperature-dependent energy, the behavior of which is studied in [14] and [19] for the asymptotic case and in [24] for the finite-time case. Besides, SC is included in the general class of Markov processes investigated in [11]. However, the convergence results in [11] and [24] require that

$$\sup_{(x,\beta)\in\Omega\times\mathbb{R}_+} \beta \left|U_\beta(x) - U(x)\right| < +\infty, \tag{48}$$

while it is assumed in [14] and [19] that there exists $a > 0$ such that

$$\sup_{(x,n)\in\Omega\times\mathbb{N}^*} n^a |U_{\beta_{n+1}}(x) - U_{\beta_n}(x)| < +\infty. \tag{49}$$

These conditions impose lower bounds on the speed of convergence of the continuation scheme which significantly limit the freedom in parameterizing the energy with temperature. Consequently, the difficulty of $(\Omega, U_{\beta_n}, \theta_{\beta_n})$ may increase too rapidly, thereby reducing — if not canceling — the benefits of continuation. Moreover, the convergence results in [11, 14, 19] involve impractical logarithmic cooling sequences.

Theorem 3 below shows that, under weak conditions, the above limitations can be overcome while allowing the communication mechanism to vary with temperature. (The proof is given in [29, 30] — it starts from the observation that SC and SA behave similarly at low temperatures in the sense that they satisfy the same large deviation principle, which allows to use the generalized SA theory developed in [8].) Given a Markov matrix q on Ω, we denote by $\mathrm{supp}(q)$ the support of q, that is, $\mathrm{supp}(q) = \{(x,y) \in \Omega^2 | q(x,y) > 0\}$, and we say that $\mathrm{supp}(q)$ is symmetric if for any $(x,y) \in \Omega^2$, $(x,y) \in \mathrm{supp}(q) \Longrightarrow (y,x) \in \mathrm{supp}(q)$. We recall that H_{c}, D and D_{M} are the critical depth, the difficulty and the "Metropolis difficulty" of (Ω, U, θ) defined in (14), (34) and (41), respectively.

Theorem 3. *Let $(\Omega, (U_\beta), (\theta_\beta))$ be an SC process with target energy landscape (Ω, U, θ) and satisfying the following assumptions:*

(A1) *θ is irreducible;*
(A2) $\mathrm{supp}(\theta)$ *is symmetric;*
(A3) $\forall x \in \Omega$, $\theta(x,x) > 0$*;*
(A4) $\mathrm{supp}(\theta_\beta) = \mathrm{supp}(\theta)$ *for β large enough.*

For any $\varepsilon > 0$ and for any $\sigma \in \mathbb{N}^*$ such that

$$\sigma > \frac{\ln(D_{\mathrm{M}}/D)}{\ln(1+\varepsilon)}, \tag{50}$$

there is a family $\{(\beta_n^{\sigma,K})_{n\in[\![1,\sigma K]\!]}\}_{K\in\mathbb{N}^*}$ *of piecewise-constant cooling sequences (*σ *denotes the number of constant-temperature stages, each of length K) such that the family of finite-time algorithms*

$$\left\{ \left(X_n^{\sigma,K}\right)_{n\in[\![1,\sigma K]\!]} = \mathrm{SC}\big(\Omega, (U_\beta), (\theta_\beta), (\beta_n^{\sigma,K})_{n\in[\![1,\sigma K]\!]}\big) \; ; \; K \in \mathbb{N}^* \right\} \tag{51}$$

satisfies

$$\lim_{K\to+\infty} -\frac{\ln \sup_{x\in\Omega} \mathsf{P}\big(X_{\sigma K}^{\sigma,K} \notin \Omega_{\mathrm{inf}} \,\big|\, X_0^{\sigma,K} = x\big)}{\ln(\sigma K)} \geqslant \frac{1}{(1+\varepsilon)D}. \tag{52}$$

These cooling sequences are of the form

$$\beta_n^{\sigma,K} = \frac{\ln K}{A} \exp\left(\frac{B}{\sigma}\left(\left\lceil \frac{n}{K} \right\rceil - 1\right)\right) \tag{53}$$

$$\textit{with} \quad \begin{cases} A > H_{\mathrm{c}}, \\ \ln(D_{\mathrm{M}}/D) < B < \sigma \ln(1+\varepsilon). \end{cases} \tag{54}$$

If (A1)–(A4) hold, then Theorem 3 gives that for any $\alpha \in (0, 1/D)$, there is a family of piecewise-constant exponential cooling sequences of the form (53) such that

$$\sup_{x\in\Omega} \mathsf{P}\big(X_{\sigma K}^{\sigma,K} \notin \Omega_{\mathrm{inf}} \,\big|\, X_0^{\sigma,K} = x\big) \leqslant (\sigma K)^{-\alpha} \tag{55}$$

for K large enough. In other words, increasing the length of the temperature stages of piecewise-constant exponential cooling makes it possible for SC to have a convergence speed exponent arbitrarily close to the optimal exponent of SA. Interestingly, the assumptions of Theorem 3 do not involve the continuation scheme $(U_\beta)_\beta$ (except for pointwise convergence to the target energy). Moreover, it is easy to construct a communication scheme $(\theta_\beta)_\beta$ satisfying (A1)–(A4). Assumptions (A1) and (A2) are standard in SA theory: the irreducibility of θ and the symmetry of its support ensure that the target energy landscape can be fully explored and that any path in this landscape can be traveled in the opposite direction (note that it is not necessary that θ be symmetric). Assumptions (A3) and (A4) mean that the limit communication mechanism can rest anywhere and that the set of possible moves is "frozen" at low temperatures.

6.3 Design Guidelines

The generation of a realization $(x_n)_n$ of a continuation chain $\mathrm{SC}(\Omega, (U_\beta), (\theta_\beta), (\beta_n))$ is the same as that of an annealing chain $\mathrm{SA}(\Omega, U, \theta, (\beta_n))$, but with U and θ

respectively replaced by U_{β_n} and θ_{β_n}. For a piecewise-constant cooling sequence $(\beta_n^{\sigma,K})_{n \in [\![1,\sigma K]\!]}$ with σ stages of length K, the construction is the following:

```
pick an initial state $x_0 \in \Omega$;
for $i = 1$ to $\sigma$ do
    set $\beta \longleftarrow \beta_{(i-1)K+1}^{\sigma,K}$;
    for $j = 1$ to $K$ do
        set $n \longleftarrow (i-1)K + j$;
        draw a state $y$ from the probability distribution $\theta_\beta(x_{n-1}, \cdot)$ on $\Omega$;
        set $x_n \longleftarrow x_{n-1}$;
        set $\delta \longleftarrow U_\beta(y) - U_\beta(x_{n-1})$;
        if $\delta \leqslant 0$ then set $x_n \longleftarrow y$;
        else set $x_n \longleftarrow y$ with probability $\exp(-\beta\delta)$;
        end(if)
    end(for)
end(for)
```

The time-complexity of SC is governed by the evaluation of the energy difference that takes place at each iteration. Let $\mathscr{T}_\beta(x,y)$ and $\mathscr{T}(x,y)$ be the time-complexities of computing $U_\beta(y) - U_\beta(x)$ and $U(y) - U(x)$, respectively. The choice of the continuation and communication schemes $(U_\beta)_\beta$ and $(\theta_\beta)_\beta$ can be guided by the objective of keeping the weighted average $\sum_{(x,y)\in\Omega^2} \theta_\beta(x,y)\, \mathscr{T}_\beta(x,y)$ of the same order as $\sum_{(x,y)\in\Omega^2} \theta(x,y)\, \mathscr{T}(x,y)$. In this case, putting aside possible updating operations at the beginning of each temperature stage, SC with piecewise-constant cooling has the same time-complexity as SA.

Ideally, $(U_\beta)_\beta$ should be designed so that the difficulty of $(\Omega, U_\beta, \theta)$ increases with increasing β. According to Theorems 1 and 2 in Section 5, a simple idea is to use a parameterized concave transform with decreasing concavity-to-increase ratio, that is, to set $U_\beta = \varphi_\beta \circ U$, where $(\varphi_\beta)_\beta$ is a family of increasing, strictly concave, twice differentiable functions such that $-\varphi_\beta''/\varphi_\beta'$ decreases as β increases. Except for this particular construction, the design of $(U_\beta)_\beta$ cannot generally be guided by the variations of $D(\Omega, U_\beta, \theta)$ with β, as estimating the difficulty of an energy landscape is intractable in most practical situations. However, it is often possible to exploit some particular characteristics of the target energy function to construct an efficient continuation scheme; example applications include image reconstruction [24, 27], where β controls the non-convexity of the energy function, inverse treatment planning in radiotherapy [31], where β controls the strength of the constraints aimed at sparing the critical tissues, and graph layout [30], where β controls the size of the ideal edge-length.

Intuitively, the communication scheme $(\theta_\beta)_\beta$ should allow balanced exploration of the state space at the beginning of the SC process, and it should favor moves towards nearby minima by the end of the SC process. A simple and efficient way to get this behavior is to design two communication matrices $\overline{\theta}$ and $\underline{\theta}$ that are respectively adapted to the high- and low-temperature regimes, and to control the probability of choosing one over the other as a function of β; that is,

$$\theta_\beta = (1-\xi(\beta))\,\overline{\theta} + \xi(\beta)\,\underline{\theta}, \tag{56}$$

where $\xi(\beta)$ is the probability of choosing $\underline{\theta}$ rather than $\overline{\theta}$ to generate a candidate solution. The control function $\xi : \mathbb{R}_+ \to [0,1]$ is monotonically increasing, and we can impose that $\lim_{\beta\to+\infty}\xi(\beta) < 1$ to place the conditions of Theorem 3 on $\overline{\theta}$; in this case, (A1)–(A4) hold if (i) $\overline{\theta}$ is irreducible, (ii) $\overline{\theta}(x,x) > 0$ for all x, (iii) $\operatorname{supp}(\overline{\theta})$ is symmetric, and (iv) $\operatorname{supp}(\underline{\theta}) \subseteq \operatorname{supp}(\overline{\theta})$. Concrete examples of using communication schemes of the type of (56) can be found in [27, 31, 30].

Another interesting possibility is *hierarchical SA*, which consists in progressively refining the exploration of the target energy landscape by operating on a hierarchy of nested approximation spaces associated to different temperature intervals. This hierarchy is defined by a sequence $(\Omega_r)_{r\in[\![1,\rho]\!]}$ of subsets of Ω such that $\emptyset \neq \Omega_1 \subset \cdots \subset \Omega_\rho = \Omega$ and by a partition of $\mathbb{R}_+$ into ρ successive intervals $I_1,\ldots,I_\rho$. For each $r \in [\![1,\rho]\!]$, the subspace Ω_r is the approximation space to be explored when the inverse temperature β is in I_r, and Ω_r is associated to an energy function $V_r : \Omega_r \to \mathbb{R}$ approximating U on Ω_r and to a communication matrix $q_r : \Omega_r^2 \to [0,1]$ adapted to the exploration of Ω_r, with the obvious requirement that $(V_\rho, q_\rho) = (U,\theta)$. A hierarchical SA process $(\Omega_r, I_r, V_r, q_r)_{r\in[\![1,\rho]\!]}$ is an SC process with continuation and communication schemes defined as follows: for all $r \in [\![1,\rho]\!]$ and for all $\beta \in I_r$, U_β is any extension of V_r to Ω, and $\theta_\beta = q_r$ on Ω_r^2 and is zero elsewhere. The hierarchical approach is interesting when the considered optimization problem lends itself to a multiscale, coarse-to-fine analysis, which is typically the case when Ω is a cartesian product space indexed by the sites of a large spatial lattice, as in image processing problems such as denoising, reconstruction and segmentation. To achieve good performance, for each $r \in [\![2,\rho]\!]$, the communication matrix q_r should be adapted to the exploration of the neighborhoods in (Ω,U,θ) that correspond to the detail difference between the states in Ω_r and their coarser representations in Ω_{r-1}.

7 Practical Tuning of the Cooling Sequence

We know from Sections 4 and 6 that exponential cooling is the best choice for both SA and SC. However, although Theorem 3 provides bounds for tuning the cooling sequence, it is generally not possible to obtain good estimates of the critical constants of the target energy landscape—at least not in a reasonable amount of computation time—, and thus the problem of choosing appropriate cooling parameters remains.

Consider an exponential cooling sequence $(\beta_n^{\sigma,K})_{n\in[\![1,\sigma K]\!]}$ with σ constant-temperature stages of length K. This sequence can be written under the form

$$\beta_n^{\sigma,K} = \beta_{\inf}\left(\frac{\beta_{\sup}}{\beta_{\inf}}\right)^{\frac{1}{\sigma-1}\left(\left\lceil\frac{n}{K}\right\rceil-1\right)}, \tag{57}$$

where $\beta_{\inf}$ and $\beta_{\sup}$ respectively denote the initial and final inverse temperatures. The horizon σK is generally fixed by the available computing resources, and setting σ is not critical, as the performance of SC is robust to the choice of σ if σ is large enough (to fix ideas, $\sigma \geqslant 100$ is adequate for most cases). This leaves us with the issue of setting $\beta_{\inf}$ and $\beta_{\sup}$, which has been addressed by many authors in the early ages of SA [18]. According to our experience, the two heuristics below yield consistently good results. We recall that Q_β is the transition matrix of SC at temperature β^{-1} (as defined by (47)) and that ν_β denotes the invariant measure of Q_β.

1. Most transitions should be accepted at the beginning of the optimization process; that is, letting $(X_n)_n$ be the homogeneous Markov chain with transition matrix $Q_{\beta_{\inf}}$, the acceptance rate

   $$\sum_{x\in\Omega} \nu_{\beta_{\inf}}(x) \sum_{y\in\Omega\setminus\{x\}} Q_{\beta_{\inf}}(x,y) \;=\; \operatorname*{plim}_{M\to+\infty} \frac{1}{M} \sum_{n=1}^{M} \mathbf{1}_{\{X_n\neq X_{n-1}\}}$$

   should be close to one. (The existence of the probability limit follows from the irreducibility of $Q_{\beta_{\inf}}$.)
2. If a global minimum $x^\dagger$ of the target energy U is reached by the end of the optimization process, then the probability to leave $x^\dagger$ by moving uphill should be negligible; that is,

   $$\sum_{z\in\Omega\,:\,U(z)>U(x^\dagger)} Q_{\beta_{\sup}}(x^\dagger,z)$$

   should be close to zero. (In practice, $x^\dagger$ must be replaced with a local minimum computed deterministically, as the ultimate goal is precisely to find a ground state of U.)

Accurate methods to estimate $\beta_{\inf}$ and $\beta_{\sup}$ according to the above criteria can be found in [28], but they are time-consuming. Besides, high accuracy is not necessary because exponential cooling is not greatly affected by excessively high initial temperatures or by excessively low final temperatures. The truth is that, as long as the horizon σK is large enough, correct orders of magnitude are satisfactory and hence fast approximate estimation methods are sufficient. In this spirit, we propose to select $\beta_{\inf}$ and $\beta_{\sup}$ so that the uphill acceptance rates (that is, the ratios of the number of accepted uphill moves to the number of proposed ones) at the beginning and at the end of the optimization process are close to some given values $\chi_{\beta_{\inf}}$ and $\chi_{\beta_{\sup}}$ such that $0 < \chi_{\beta_{\sup}} \ll \chi_{\beta_{\inf}} < 1$. For this purpose, the initial energy landscape $(\Omega, U_{\beta_{\inf}}, \theta_{\beta_{\inf}})$ is approximated by the infinite-temperature energy landscape (Ω, U_0, θ_0), and the final energy landscape $(\Omega, U_{\beta_{\sup}}, \theta_{\beta_{\sup}})$ is approximated by the target energy landscape (Ω, U, θ). The procedures are the following.

1. The estimation of $\beta_{\inf}$ uses the Markov chain $(X_n)_n$ defined by the communication matrix θ_0: given $M \in \mathbb{N}^*$, generate a finite-time realization $(x_n)_n$ of $(X_n)_n$ with exactly M uphill moves with respect to U_0 (that is, M pairs (x_{n_k}, x_{n_k+1}) of successive states such that $U_0(x_{n_k}) < U_0(x_{n_k+1})$, $k \in [\![1,M]\!]$), and set $\beta_{\inf}$ to be the solution of

$$\sum_{k=1}^{M} \exp\bigl(-\beta(U_0(x_{n_k+1}) - U_0(x_{n_k}))\bigr) \;=\; M\chi_{\beta_{\mathrm{inf}}}, \tag{58}$$

which can be determined by any standard root-finding method.

2. Similarly, β_{sup} is estimated from a realization $(y_n)_n$ of the Markov chain with transition matrix θ by considering the M first uphill moves (y_{n_k}, y_{n_k+1}) with respect to the target energy U; that is, β_{sup} is set to be the solution of

$$\sum_{k=1}^{M} \exp\bigl(-\beta(U(y_{n_k+1}) - U(y_{n_k}))\bigr) \;=\; M\chi_{\beta_{\mathrm{sup}}}. \tag{59}$$

Looking only for estimates with correct orders of magnitude gives some latitude in choosing $\chi_{\beta_{\mathrm{inf}}}$ and $\chi_{\beta_{\mathrm{sup}}}$: taking $\chi_{\beta_{\mathrm{inf}}} \in [0.6, 0.9]$ and $\chi_{\beta_{\mathrm{sup}}} \in [10^{-4}, 10^{-3}]$ gives exponential cooling schedules with similar performance independently of the application. The number M of considered uphill moves can be set in accordance with the size of the optimization problem; for instance, choosing M of the order of $100d$ is suitable for the case where Ω is a cartesian product space included in $\mathbb{R}^d$.

8 Conclusion

Despite its popularity, simulated annealing (SA) remains largely criticized for its slow convergence. This criticism is fully justified if we stick to early convergence results which impose unfeasible logarithmic cooling schedules. In practice, one usually takes liberties with the design of SA algorithms at the expense of losing global convergence guarantees, and it is commonly admitted that SA implementations are suboptimal.

Our objective was to emphasize advanced theoretical developments and design guidelines for annealing-type algorithms. In particular, we have seen that exponential cooling makes it possible for the probability of failure to decrease to zero with a speed exponent arbitrarily close to the optimal exponent, and that inexpensive acceleration techniques such as restriction of the state space, transformation of the state space, and concave distortion of the energy function can increase the performance of SA while not altering its global convergence properties. Even more importantly, we have shown that increasing the flexibility by allowing the communication mechanism and the energy function to vary with temperature is theoretically grounded. This generalization of SA, called stochastic continuation, has global convergence properties similar to that of standard SA under weak assumptions on the communication mechanism and independently of the speed of convergence of the energy towards the target objective function. Ultimately, then, the advances in SA theory presented in this paper make annealing-type algorithms attractive for a wide range of difficult optimization problems.

Acknowledgements. This work was partly supported by the French National Research Agency under grant ANR-09-BLAN-0372-01.

References

1. Discussion on the meeting on the Gibbs sampler and other Markov chain Monte Carlo methods. J. Roy. Statist. Soc. Ser. B 55(1), 53–102 (1993)
2. Azencott, R.: A common large deviations mathematical framework for sequential annealing and parallel annealing. In: Azencott, R. (ed.) Simulated Annealing: Parallelization Techniques, pp. 11–23. Wiley, New York (1992)
3. Azencott, R.: Sequential simulated annealing: speed of convergence and acceleration techniques. In: Azencott, R. (ed.) Simulated Annealing: Parallelization Techniques, pp. 1–10. Wiley, New York (1992)
4. Blake, A., Zisserman, A.: Visual reconstruction. The MIT Press (1987)
5. Catoni, O.: Large deviations and cooling schedules for the simulated annealing algorithm. C. R. Acad. Sci. Paris Sér. I Math. 307, 535–538 (1988) (in French)
6. Catoni, O.: Rough large deviation estimates for simulated annealing: application to exponential schedules. Ann. Probab. 20(3), 1109–1146 (1992)
7. Catoni, O.: Metropolis, simulated annealing, and iterated energy transformation algorithms: theory and experiments. J. Complexity 12(4), 595–623 (1996)
8. Catoni, O.: Simulated annealing algorithms and Markov chains with rare transitions. In: Séminaire de Probabilités XXXIII. Lecture Notes in Math., vol. 1709, pp. 69–119. Springer, New York (1999)
9. Chiang, T.S., Chow, Y.: On the convergence rate of annealing processes. SIAM J. Control Optim. 26(6), 1455–1470 (1988)
10. Cohn, H., Fielding, M.: Simulated annealing: searching for an optimal temperature schedule. SIAM J. Optim. 9(3), 779–802 (1999)
11. Del Moral, P., Miclo, L.: On the convergence and applications of generalized simulated annealing. SIAM J. Control Optim. 37(4), 1222–1250 (1999)
12. Desai, M.: Some results characterizing the finite time behaviour of the simulated annealing algorithm. Sādhanā 24(4-5), 317–337 (1999)
13. Fielding, M.: Simulated annealing with an optimal fixed temperature. SIAM J. Optim. 11(2), 289–307 (2000)
14. Frigerio, A., Grillo, G.: Simulated annealing with time-dependent energy function. Math. Z. 213, 97–116 (1993)
15. Gidas, B.: Nonstationary Markov chains and convergence of the annealing algorithm. J. Statist. Phys. 39(1/2), 73–131 (1985)
16. Hajek, B.: Cooling schedules for optimal annealing. Math. Oper. Res. 13(2), 311–329 (1988)
17. Johnson, A., Jacobson, S.: On the convergence of generalized hill climbing algorithms. Discrete Appl. Math. 119(1-2), 37–57 (2002)
18. van Laarhoven, P.J.M., Aarts, E.H.L.: Simulated annealing: theory and practice. D. Reidel Publishing Company (1987)
19. Löwe, M.: Simulated annealing with time-dependent energy function via Sobolev inequalities. Stochastic Process. Appl. 63(2), 221–233 (1996)
20. Nielsen, M.: Graduated nonconvexity by functional focusing. IEEE Trans. Pattern Anal. Machine Intell. 19(5), 521–525 (1997)
21. Nikolova, M.: Markovian reconstruction using a GNC approach. IEEE Trans. Image Process. 8(9), 1204–1220 (1999)
22. Orosz, J., Jacobson, S.: Analysis of static simulated annealing algorithms. J. Optim. Theory Appl. 115(1), 165–182 (2002)
23. Orosz, J., Jacobson, S.: Finite-time performance analysis of static simulated annealing algorithms. Comput. Optim. Appl. 21(1), 21–53 (2002)

24. Robini, M.C., Lachal, A., Magnin, I.E.: A stochastic continuation approach to piecewise constant reconstruction. IEEE Trans. Image Process. 16(10), 2576–2589 (2007)
25. Robini, M.C., Magnin, I.E.: 3-D reconstruction from a few radiographs using the Metropolis dynamics with annealing. In: Proc. IEEE Int. Conf. Image Processing, Kobe, Japan, vol. 3, pp. 876–880 (1999)
26. Robini, M.C., Magnin, I.E.: Stochastic nonlinear image restoration using the wavelet transform. IEEE Trans. Image Process. 12(8), 890–905 (2003)
27. Robini, M.C., Magnin, I.E.: Optimization by stochastic continuation. SIAM J. Imaging Sci. 3(4), 1096–1121 (2010)
28. Robini, M.C., Rastello, T., Magnin, I.E.: Simulated annealing, acceleration techniques and image restoration. IEEE Trans. Image Process. 8(10), 1374–1387 (1999)
29. Robini, M.C., Reissman, P.J.: On simulated annealing with temperature-dependent energy and temperature-dependent communication. Statist. Probab. Lett. 81(8), 915–920 (2011)
30. Robini, M.C., Reissman, P.J.: From simulated annealing to stochastic continuation: a new trend in combinatorial optimization. J. Global Optim. (to appear, 2012)
31. Robini, M.C., Smekens, F., Sixou, B.: Optimal inverse treatment planning by stochastic continuation. In: Proc. 8th IEEE Int. Symp. Biomedical Imaging, Chicago, IL, pp. 1792–1796 (2011)
32. Schuur, P.: Classification of acceptance criteria for the simulated annealing algorithm. Math. Oper. Res. 22(2), 266–275 (1997)