SIAM J. IMAGING SCIENCES
Vol. 3, No. 4, pp. 1096–1121

Optimization by Stochastic Continuation*

Marc C. Robini† and Isabelle E. Magnin†

Abstract. Simulated annealing (SA) and deterministic continuation are well-known generic approaches to global optimization. Deterministic continuation is computationally attractive but produces suboptimal solutions, whereas SA is asymptotically optimal but converges very slowly. In this paper, we introduce a new class of hybrid algorithms which combines the theoretical advantages of SA with the practical advantages of deterministic continuation. We call this class of algorithms stochastic continuation (SC). In a nutshell, SC is a variation of SA in which both the energy function and the communication mechanism are allowed to be time-dependent. We first prove that SC inherits the convergence properties of generalized SA under weak assumptions. Then, we show that SC can be successfully applied to optimization issues raised by the Bayesian approach to signal reconstruction. The considered class of energy functions arises in maximum a posteriori estimation with a Markov random field prior. The associated minimization task is NP-hard and beyond the scope of popular methods such as loopy belief propagation, tree-reweighted message passing, and graph cuts and its extensions. We perform numerical experiments in the context of three-dimensional reconstruction from a very limited number of projections; our results show that SC can substantially outperform both deterministic continuation and SA.

Key words. optimization, simulated annealing, Markov chain Monte Carlo, continuation, signal reconstruction, inverse problems

AMS subject classifications. 82C80, 65C05, 60J10, 90C27, 49N45, 68U99, 94A08, 94A12

DOI. 10.1137/090756181

1. Introduction.

1.1. Background.

Many computer vision problems and signal processing tasks involve global optimization. Typical examples include image restoration [52], image segmentation [32], boundary detection [22], stereo matching [59], motion estimation [66], texture analysis [25], sparse approximation [13], filter design [10], and network optimization [2]. The challenge is to overcome the multimodality of complex cost functions, which often traps algorithms in poor local minima.

Promising specific optimization approaches have emerged in the last few years; the most popular are graph cuts [6] and its extensions based on quadratic pseudo-Boolean optimization (QPBO) [56, 34], loopy belief propagation (LBP) [63], and tree-reweighted message passing (TRW) [35]. These techniques are usually confined to energies of the form

$$x \in \Lambda^K \longmapsto \sum_{k \in [\![1,K]\!]} \mathcal{J}_k(x_k) + \sum_{\{k,l\} \in \mathcal{E}} \phi_{\{k,l\}}(x_k, x_l), \tag{1.1}$$

*Received by the editors April 16, 2009; accepted for publication (in revised form) July 13, 2010; published electronically December 21, 2010. This work was supported by the French National Research Agency under grant ANR-09-BLAN-0372-01.

http://www.siam.org/journals/siims/3-4/75618.html

†CREATIS (CNRS research unit UMR 5220 and INSERM research unit U630), INSA-Lyon, 7 av. Jean Capelle, 69621 Villeurbanne cedex, France (marc.robini@creatis.insa-lyon.fr, isabelle.magnin@creatis.insa-lyon.fr).

where Λ is a finite set of labels, the $\mathcal{J}_k$'s are unary data-likelihood functions, and the $\phi_{\{k,l\}}$'s are pairwise interaction potentials ($[\![1, K]\!]$ is a shorthand notation for $\{1, \ldots, K\}$, and $\mathcal{E}$ is a collection of 2-subsets of $[\![1, K]\!]$). Such functions typically arise in maximum a posteriori (MAP) estimation with a Markov random field (MRF) prior (see, e.g., [21, 39]). An extensive three-way comparison of graph cuts, LBP, and TRW is given in [60], where it is observed that graph cuts and TRW produce consistently high-quality results and perform better than LBP. The graph cuts method has the most interesting convergence properties among the three; it finds local minima with respect to large moves in the state space and hence generally produces near-optimal solutions. However, as shown in Appendix A, graph cuts do not apply to fundamental optimization problems such as those associated with signal reconstruction in a MAP-MRF framework, even if the corresponding energy is of type (1.1). (Common examples include image restoration and two- or three-dimensional reconstruction from line-integral projection.) The reason for this is that graph cuts are limited to energies whose pairwise interaction potentials satisfy

$$\forall \{a, b, c\} \subset \Lambda, \qquad \phi_{\{k,l\}}(a, a) + \phi_{\{k,l\}}(b, c) \leqslant \phi_{\{k,l\}}(b, a) + \phi_{\{k,l\}}(a, c)\,.$$

Such potentials are said to be *submodular*. A first approach to dealing with nonsubmodular terms is to "truncate" them, as proposed in [57], but the experiments in [36] and [56] show that this technique performs well only when the *nonsubmodularity ratio* (i.e., the ratio of the number of nonsubmodular to submodular terms) is very small. In the binary-label case, the method of choice is the QPBO procedure from [28] introduced in computer vision by Kolmogorov and Rother [36]. Yet the performance of QPBO also decreases with increasing nonsubmodularity ratio, although less rapidly than truncation. In particular, we show in Appendix B that, in the context of signal reconstruction, the behavior of QPBO is governed by the ratio of the number of pair-site cliques in the data-likelihood to the number in the prior. Our argument is substantiated by the binary image restoration experiments reported in [56], and our conclusion is that QPBO is not suitable for reconstruction problems in which the neighborhood system associated with the data-likelihood is larger than the neighborhood system in the prior. In the multilabel case, the extensions of QPBO come in two forms. The first approach is to use QPBO within the α-expansion procedure of Boykov (see [6]), as proposed in [49] and [56], and the second approach is to reduce the original multilabel energy to a function of binary variables to be minimized by QPBO [34]. It stands to reason that both methods suffer from the limitations of QPBO: their performance decreases with increasing nonsubmodularity ratio and increasing strength of the nonsubmodular terms, and the experiments in [34] indicate decreasing performance with increasing number of labels and with increasing nonconvexity of the interaction potentials. Ultimately, then, the need for efficient general-purpose global optimization methods is crucial.

Two well-known generic optimization approaches are simulated annealing (SA), pioneered in [33], and deterministic continuation, examples of which are mean-field annealing [19] and graduated nonconvexity (GNC) [4]. On the one hand, SA is asymptotically optimal [24, 27, 11] but is generally reported to converge slowly, and on the other hand, deterministic continuation has reasonable computational load but is suboptimal. In practice, deterministic continuation is preferred whenever possible; it is not so much that annealing is slow, but rather that SA is commonly dismissed based on an early study by Blake [3] which demonstrates the

superiority of GNC over the annealing approach of Geman and Geman [24] in the context of signal denoising. The truth is that carefully designed annealing algorithms produce very good results for a wide class of reconstruction problems (clear-cut examples can be found in [53, 52]), although inappropriate design of SA still appears in recent work; for instance, Nikolova et al. [48] claim that SA does not work for image deconvolution regardless of the successful results in [23, 53] (obtained for the same cost function) and mention several days of computation time for denoising a 64×64 gray-level image, whereas our annealing algorithm in [53] takes about 10 minutes on a standard PC to perform this task. Yet, despite various theoretical and practical improvements over the last two decades, SA is generally much slower than deterministic methods, and efficient acceleration techniques are welcome.

1.2. Contributions of this paper.

1.2.1. Stochastic continuation. A promising idea to speed up annealing was recently developed in [51], where it is shown that incorporating a time-dependent energy sequence into SA can substantially increase performance. More precisely, the resulting class of algorithms, called stochastic continuation (SC), inherits the finite-time convergence properties of generalized simulated annealing (GSA) established in [12], provided that the energy sequence is continuous with respect to temperature and converges fast enough to the target cost function. Yet these conditions are sometimes restrictive, and the results in [51] do not include the possibility of letting the communication mechanism vary with time. Our first contribution is to show that the restrictions on the energy sequence can be removed (the only remaining condition being pointwise convergence to the target), and that it is as possible to use time-dependent communication to facilitate the exploration of the state space. This great flexibility in the design of annealing-type algorithms is made possible by the theoretical results in [8] and opens new horizons to solving difficult optimization problems in the computer vision and signal processing fields.

Both SA and SC belong to the class of Markov chain Monte Carlo (MCMC) methods; in simple terms, SA is a speed-up technique for homogeneous MCMC optimization (see [7] for a theoretical justification), and SC is an extension of SA that allows further convergence improvement. In this paper, we limit ourselves to the case of Metropolis sampling, but the SC approach may be extended to the more general Metropolis–Hastings dynamics [29], which includes well-known MCMC samplers such as the Gibbs and the Swendsen–Wang dynamics [24, 58]. We will also assume that the state space is finite, because the existing results on MCMC optimization on general state spaces—especially continuous domains—are not flexible enough to study the behavior of SC algorithms: first, none allows the energy to be time-dependent; second, they either require logarithmic cooling [26, 20, 41, 42] or impose too restrictive conditions on the communication mechanism [1, 65]; and third, none provides relevant information on the convergence speed. We might add that the finite state space assumption is not a problem in most practical situations, since it is generally possible to increase the number of states to achieve the desired level of accuracy.

1.2.2. MAP-MRF reconstruction. The second contribution of this paper is the application of SC to optimization issues raised by the MAP-MRF approach to the classical inverse problem of estimating an unknown one- or multidimensional piecewise-constant signal $x^\star \in \mathbb{R}^K$

given data of the form

$$d = \mathcal{H}(x^\star) + e_\eta,$$

where $\mathcal{H} : \mathbb{R}^K \to \mathbb{R}^{K'}$ is a linear map that models the acquisition process and where the noise term $e_\eta \in \mathbb{R}^{K'}$ is a realization of a random vector η. The solution is defined as any global minimum of an energy function

$$U : x \in \Lambda^K \longmapsto \mathcal{J}(x) + \lambda\, \Phi(x), \tag{1.2}$$

where $\Lambda \subset \mathbb{R}$, $\mathcal{J} : \mathbb{R}^K \to \mathbb{R}$ is a data-likelihood function, $\Phi : \mathbb{R}^K \to \mathbb{R}$ is a Gibbs energy favoring solutions in agreement with some a priori knowledge about the true signal $x^\star$, and the parameter $\lambda \in \mathbb{R}_+^*$ governs the trade-off between $\mathcal{J}$ and Φ [16, 21, 39]. The prior term is given by

$$\Phi(x) = \sum_{i \in [\![1,M]\!]} \phi\big(\|\mathcal{D}_i(x)\|\big), \tag{1.3}$$

where the *potential function* $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is increasing, $\|\cdot\|$ is the ℓ_2-norm, and each $\mathcal{D}_i$ is a linear map from $\mathbb{R}^K$ to $\mathbb{R}^{K_i}$. Concrete examples can be found in [23, 40, 9, 15, 53, 45, 55]. Most of the time, the $\mathcal{D}_i$'s are either first-order difference operators ($K_i = 1$) or discrete approximations to the gradient operator ($K_i = r$ in the r-dimensional case); more sophisticated possibilities include second- or third-order differences [23, 52] and first-order differences in the wavelet domain [52, 54]. The data-likelihood term is generally either of the form

$$\mathcal{J}(x) = \|\mathcal{H}(x) - d\|_\eta^2, \tag{1.4}$$

where $\|\cdot\|_\eta$ is the ℓ_2-norm weighted by the inverse covariance matrix of η (that is, η is assumed to be independent from $x^\star$ and to follow a zero-mean multivariate normal distribution), or of the form

$$\mathcal{J}(x) = \sum_{k \in [\![1,K']\!]} \psi_k\big(\mathcal{H}_k(x) - d_k\big), \tag{1.5}$$

where the ψ_k's are even functions increasing on $\mathbb{R}_+$. An important particular case of (1.5) is the ℓ_1 data-fidelity term obtained by setting $\psi_k(t) = |t|$ for all k, which is well adapted to data corrupted with outliers or with impulsive noise [45, 55].

When $\Lambda = \mathbb{R}$, there exist efficient algorithms for finding the global minimum of U when U is C^1 and strictly convex [15], which presupposes that both $\mathcal{J}$ and ϕ are C^1 and convex and that $\phi'(0) = 0$. When Λ is finite, exact optimization can be achieved in the special case where $\mathcal{J}$ is of the form (1.4) with $\mathcal{H}$ being the identity on $\mathbb{R}^K$, the $\mathcal{D}_i$'s are first-order differences, and ϕ is convex [30]. If $\mathcal{J}$ is of the form (1.1) (as is (1.4)) and if the $\mathcal{D}_i$'s are either canonical projections or first-order differences, then it is possible to look for near-optimal solutions by multilabel QPBO [49, 56, 34], provided that the adjacency graph associated with the data-likelihood is sparsely connected (in accordance with our discussion about QPBO in section 1.1). Multilabel QPBO can also be used in conjunction with appropriate clique reduction

techniques [5, 31] if $\mathcal{J}$ is of the form (1.5) or if the $\mathcal{D}_i$'s are more sophisticated than first-order differences. Still, deterministic continuation and stochastic optimization are more viable options when the adjacency graph of the likelihood is not sparsely connected.

We focus on the case where ϕ is the "0-1" function, that is,

$$\phi(t) = \begin{cases} 1 & \text{if } t > 0, \\ 0 & \text{if } t = 0. \end{cases} \tag{1.6}$$

The reason for this choice is twofold. First, the 0-1 potential function is particularly appropriate for the recovery of piecewise-constant signals (we refer to [46] for a comprehensive treatment). Second, the corresponding optimization problem is challenging for both GNC and SA, which makes it a good test case for demonstrating the usefulness of our approach. In fact, both GNC and SA produce good results when using more standard nonconvex potentials such as the truncated-quadratic function $\phi(t) = \min(t^2, 1)$ or the "less friendly" concave function $\phi(t) = t/(1+t)$ [23, 47, 44, 53, 48]. In the case of the 0-1 function, however, the energy U is complex enough to mislead GNC tracking processes and has very narrow valleys in which SA gets easily trapped. The associated minimization task is a generalization of the metric labeling problem with the Potts model, which is NP-hard [6], and it is closely related to minimum description length estimation: U is identical to the Leclerc model [38] when $\mathcal{J}$ is of the form (1.4) and the $\mathcal{D}_i$'s are first-order differences. (The GNC sequence proposed in [38] is experimentally compared with SA and SC in section 4.)

1.3. Outline. This paper is organized as follows. In section 2, we review some of the GSA theory, and we provide our main result about the convergence of SC. Section 3 is devoted to the application of SC to the class of optimization problems described above. Experimental results are presented in section 4, where we illustrate the potential benefits of SC over deterministic continuation and SA in the context of three-dimensional reconstruction from a very limited number of projections. Concluding remarks are given in section 5.

2. Optimization by stochastic continuation.

Let E be a finite state space, and let U be an arbitrary real-valued function defined on E. We denote the ground state energy $\min_{x\in E} U(x)$ by $U_{\min}$, and we let $E_{\min}$ be the set of global minima of U, that is, $E_{\min} = \{x \in E \,|\, U(x) = U_{\min}\}$. The objective here is to find a state x that either belongs to $E_{\min}$ or is such that $U(x)$ is as close as possible to $U_{\min}$.

2.1. Foundations: Generalized simulated annealing. GSA theory was originally developed to study parallel annealing [61]; it covers other stochastic optimization processes including SA with time-dependent energy function [51] and some genetic and evolutionary algorithms [14, 17]. A GSA process on E is a 3-tuple $(E, (Q_\beta)_\beta, V)$ defined by a family $(Q_\beta : E^2 \to [0,1])_{\beta\in\mathbb{R}_+}$ of Markov matrices having rare transitions with rate function $V : E^2 \to \mathbb{R}_+ \cup \{+\infty\}$, that is,

$$\forall (x,y) \in E^2, \qquad \lim_{\beta\to+\infty} \frac{-\ln Q_\beta(x,y)}{\beta} = V(x,y) \tag{2.1}$$

with the convention that $\ln 0 = -\infty$. The rate function is assumed to be *irreducible* in the sense that for any $(x,y) \in E^2$ there exists a *V-admissible path* from x to y, that is, a path

$(\gamma_i)_{i=0}^m$ such that $\gamma_0 = x$, $\gamma_m = y$, and $V(\gamma_{i-1}, \gamma_i) < +\infty$ for all $i \in [\![1, m]\!]$. Given such a family $(Q_\beta)_\beta$, a GSA algorithm is a Markov chain $(X_n)_{n\in\mathbb{N}}$ on E with transitions

$$\mathrm{P}(X_n = y \,|\, X_{n-1} = x) \;=\; Q_{\beta_n}(x, y),$$

where the so-called *cooling schedule* $(\beta_n)_{n\in\mathbb{N}^*}$ is a divergent, nondecreasing, positive real sequence.

A simple example is provided by (standard) SA. An SA algorithm with energy function U has transitions defined by

$$Q_\beta^{\mathrm{SA}}(x, y) \;=\; \begin{cases} q(x, y) \exp\bigl(-\beta(U(y) - U(x))^+\bigr) & \text{if } y \neq x, \\ 1 - \sum_{z\neq x} Q_\beta^{\mathrm{SA}}(x, z) & \text{if } y = x, \end{cases} \tag{2.2}$$

where $a^+ := \max\{a, 0\}$ and q is a Markov matrix on E which specifies how to generate a new candidate solution from the current one. The matrix q is called the *communication mechanism* and is assumed to have symmetric support and to be irreducible; in other words, for all $(x, y) \in E^2$, $q(x, y) > 0 \implies q(y, x) > 0$ and there exists a *q-admissible path* from x to y, that is, a path $(\gamma_i)_{i=0}^m$ such that $\gamma_0 = x$, $\gamma_m = y$, and $q(\gamma_{i-1}, \gamma_i) > 0$ for all $i \in [\![1, m]\!]$. It is shown in Appendix C that $(Q_\beta^{\mathrm{SA}})_{\beta\in\mathbb{R}_+}$ has rare transitions with rate function

$$V^{\mathrm{SA}}(x, y) \;=\; \begin{cases} (U(y) - U(x))^+ & \text{if } Q_\beta^{\mathrm{SA}}(x, y) > 0 \quad \forall \beta > 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{2.3}$$

This makes $(E, (Q_\beta^{\mathrm{SA}})_\beta, V^{\mathrm{SA}})$ a GSA process, since the irreducibility of q implies the irreducibility of V^{SA}.

A basic condition for a GSA algorithm to find the ground states of U is that its rate function V is *induced by* U; that is,

$$\forall (x, y) \in E^2, \qquad U(x) + V(x, y) \;=\; U(y) + V(y, x).$$

This is, for instance, the case of (2.3): since q has symmetric support, $V^{\mathrm{SA}}(x, y) < +\infty$ if and only if $V^{\mathrm{SA}}(y, x) < +\infty$, and thus if $V^{\mathrm{SA}}(x, y) < +\infty$, then

$$V^{\mathrm{SA}}(x, y) - V^{\mathrm{SA}}(y, x) \;=\; (U(y) - U(x))^+ - (U(x) - U(y))^+ \;=\; U(y) - U(x).$$

It is shown in [8] that if V is induced by U, then for β large enough, the invariant probability measure μ_β of Q_β satisfies

$$\forall x \in E, \qquad \lim_{\beta\to+\infty} \frac{-\ln \mu_\beta(x)}{\beta} \;=\; U(x) - U_{\min}\,. \tag{2.4}$$

An immediate consequence of (2.4) is that as β goes to infinity, μ_β gives a strictly positive mass to $E_{\min}$ and tends to zero elsewhere. Hence the idea that if $(\beta_n)_n$ does not increase too rapidly, then the law of X_n should be close enough to μ_{β_n} to achieve asymptotic optimality. As a matter of fact, the convergence measure

$$\mathcal{M}(n) \;=\; \max_{x\in E} \mathrm{P}\bigl(X_n \notin E_{\min} \,\big|\, X_0 = x\bigr) \tag{2.5}$$

cannot decrease faster than some optimal exponent of n^{-1}. Indeed, from [62],

$$\lim_{n\to+\infty} \sup_{\beta_n\geqslant\cdots\geqslant\beta_1>0} \frac{-\ln \mathcal{M}(n)}{\ln n} \leqslant \frac{1}{D},$$

where $D \in \mathbb{R}_+^*$ is called the *difficulty of the energy landscape*, the precise definition of which is given in section 2.2. The constant D is sharp since for any $\alpha < 1/D$ it is possible to construct finite cooling sequences $(\beta_n^N)_{1\leqslant n\leqslant N}$ so that $\mathcal{M}(N) \leqslant N^{-\alpha}$ for N large enough. In other words, there exist cooling schedules such that the convergence speed exponent is arbitrarily close to the optimal exponent $1/D$. Theorem 2.1 states this formally.

Theorem 2.1 (see [8]). *Let $(E, (Q_\beta)_\beta, V)$ be a GSA process with rate function induced by U. For any $\varepsilon \in \mathbb{R}_+^*$ there exist finite cooling schedules $(\beta_n^N)_{1\leqslant n\leqslant N}$ such that*

$$\liminf_{N\to+\infty} \frac{-\ln \mathcal{M}(N)}{\ln N} \geqslant \frac{1}{(1+\varepsilon)D}.$$

These schedules are piecewise-constant exponential sequences of the form

$$\beta_n^N = \beta_{\min}\left(\frac{\beta_{\max}}{\beta_{\min}}\right)^{\frac{1}{\nu-1}\left(\left\lceil \frac{\nu}{N} n\right\rceil - 1\right)}, \tag{2.6}$$

where ν is the number of temperature stages and where the minimum and maximum inverse temperatures $\beta_{\min}$ and $\beta_{\max}$ are functions of the horizon N.

Remark 1. If (2.1) is replaced by the stronger condition

$$\begin{aligned} &\exists a \in (1,+\infty),\ \ \forall (x,y) \in E^2,\ \ \forall \beta \in \mathbb{R}_+^*, \\ &\qquad a^{-1}\mathrm{e}^{-\beta V(x,y)} \leqslant Q_\beta(x,y) \leqslant a\,\mathrm{e}^{-\beta V(x,y)} \end{aligned} \tag{2.7}$$

with the convention that $\mathrm{e}^{-\beta(+\infty)} = 0$, then a theorem by Cot and Catoni [12] shows that there exist cooling schedules of the form (2.6) such that $\mathcal{M}(N)$ is asymptotically equivalent to $N^{-1/D}$ in the logarithmic scale; that is,

$$\lim_{N\to+\infty} \frac{-\ln \mathcal{M}(N)}{\ln N} = \frac{1}{D}. \tag{2.8}$$

2.2. Difficulty of the energy landscape. The 3-tuple (E, U, V) is called an *energy landscape* if the rate function V is induced by U. Assuming this is the case, the *difficulty* of (E, U, V) is given by

$$D(E,U,V) = \max_{x\in E\setminus E_{\min}} \min_{y\in E_{\min}} \frac{H(x,y)-U(x)}{U(x)-U_{\min}}$$

$$\text{with} \qquad H(x,y) = \min_{(\gamma_i)_{i=0}^m \in \Gamma_{xy}^V} \max_{i\in[\![1,m]\!]} \big(U(\gamma_{i-1}) + V(\gamma_{i-1},\gamma_i)\big), \tag{2.9}$$

where Γ_{xy}^V denotes the set of V-admissible paths from x to y.

The particular case of SA (2.2) gives some intuition about this definition. In the SA framework, the energy landscape is the 3-tuple (E, U, q), and a state $x \in E$ is called a local

minimum of (E,U,q) if $U(x) \leqslant U(y)$ for all $y \in E$ such that $q(x,y) > 0$. The energy level separating two states x and y is given by

$$h(x,y) = \min_{(\gamma_i)_{i=0}^m \in \Gamma_{xy}^q} \max_{i \in [\![0,m]\!]} U(\gamma_i), \tag{2.10}$$

where Γ_{xy}^q is the set of q-admissible paths from x to y, and we define the *depth* $\delta(x)$ of x to be the magnitude of the energy barrier separating x from a ground state:

$$\delta(x) = \min_{y \in E_{\min}} h(x,y) - U(x). \tag{2.11}$$

The difficulty of the energy landscape (E,U,q) is

$$D^{\mathrm{SA}}(E,U,q) = \max_{x \in E \setminus E_{\min}} \frac{\delta(x)}{U(x) - U_{\min}}. \tag{2.12}$$

It can be checked that the maximum in (2.12) can be taken over $E_{\mathrm{loc}} \setminus E_{\min}$, where E_{loc} denotes the set of local minima of (E,U,q). Therefore, simply speaking, D^{SA} is the maximum ratio of the depth of the nonglobal local minima to their energy level above the ground state energy.

2.3. Stochastic continuation. In a nutshell, SC is a variant of SA where both the energy function and the communication mechanism are allowed to be time-dependent. More precisely, given a cooling schedule $(\beta_n)_n$, we call an SC algorithm a Markov chain $(X_n)_{n \in \mathbb{N}}$ on E with transitions $\mathrm{P}(X_n = y \,|\, X_{n-1} = x) = Q_{\beta_n}^{\mathrm{SC}}(x,y)$ defined by

$$Q_\beta^{\mathrm{SC}}(x,y) = \begin{cases} q_\beta(x,y) \exp\bigl(-\beta(U_\beta(y) - U_\beta(x))^+\bigr) & \text{if } y \neq x, \\ 1 - \sum_{z \neq x} Q_\beta^{\mathrm{SC}}(x,z) & \text{if } y = x, \end{cases} \tag{2.13}$$

where $(q_\beta)_{\beta \in \mathbb{R}_+}$ is a family of Markov matrices on E and $(U_\beta)_{\beta \in \mathbb{R}_+}$ is a family of real-valued functions on E—we will use the notation $(E,(q_\beta),(U_\beta),(\beta_n))$ for short. Because the objective is to minimize U, this definition makes sense only if $\lim_{\beta \to +\infty} U_\beta(x) = U(x)$ for all $x \in E$. In this case, since E is finite, the global minima of U_β belong to $E_{\min}$ for β sufficiently large.

SC includes SA with time-dependent energy function, the convergence of which has been studied in [18] and [43], and more recently in [51]. Besides the fact that these papers assume a time-invariant communication mechanism, the results in [18, 43] involve impractical logarithmic cooling schedules, while it is assumed in [51] that

$$\sup_{(x,\beta) \in E \times \mathbb{R}_+} \beta \bigl| U_\beta(x) - U(x) \bigr| < +\infty, \tag{2.14}$$

which limits the freedom in designing the sequence $(U_\beta)_\beta$. Theorem 2.2 below allows us to overcome these limitations; it gives simple sufficient conditions for SC to inherit the convergence properties of GSA. (Given a Markov matrix q on E, we denote by $\mathrm{supp}(q)$ the support of q, that is, $\mathrm{supp}(q) = \{(x,y) \in E^2 \,|\, q(x,y) > 0\}$, and we say that $\mathrm{supp}(q)$ is symmetric if for any $(x,y) \in E^2$, $(x,y) \in \mathrm{supp}(q) \implies (y,x) \in \mathrm{supp}(q)$.)

Theorem 2.2. *Let $(E,(q_\beta),(U_\beta),(\beta_n))$ be an SC algorithm. Assume that*

(i) *for all* $x \in E$, $\lim_{\beta\to+\infty} U_\beta(x) = U(x)$,
(ii) *for all* $(x,y) \in E^2$, $\lim_{\beta\to+\infty} q_\beta(x,y) = q^\star(x,y)$,

where $q^\star$ *is a Markov matrix on* E *satisfying the following conditions:*

(iii) $q^\star$ *is irreducible,*
(iv) $\mathrm{supp}(q^\star)$ *is symmetric,*
(v) *for all* $x \in E$, $q^\star(x,x) > 0$,
(vi) *for all* $(x,y) \in E^2$, $q^\star(x,y) = 0 \implies \lim_{\beta\to+\infty} -\beta^{-1} \ln q_\beta(x,y) = +\infty$.

Then $(E, (q_\beta), (U_\beta), (\beta_n))$ *is a GSA algorithm with rate function induced by* U. *This rate function is given by*

$$V^{\mathrm{SC}}(x,y) = \begin{cases} (U(y) - U(x))^+ & \text{if } q^\star(x,y) > 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Proof. The proof is given in Appendix D. ∎

Let $(x,y) \in E^2$. For any V^{SC}-admissible path $(\gamma_i)_{i=0}^m$ from x to y, we have

$$U(\gamma_{i-1}) + V^{\mathrm{SC}}(\gamma_{i-1},\gamma_i) = U(\gamma_{i-1}) + (U(\gamma_i) - U(\gamma_{i-1}))^+ = \max\{U(\gamma_{i-1}), U(\gamma_i)\}$$

for all $i \in [\![1,m]\!]$. Further, since $V^{\mathrm{SC}}(x,y) < +\infty$ if and only if $q^\star(x,y) > 0$, the set of V^{SC}-admissible paths from x to y is the same as the set of $q^\star$-admissible paths from x to y. It follows that the function H from (2.9) associated with the energy landscape (E,U,V^{SC}) reduces to the function h from (2.10) associated with $(E,U,q^\star)$, and thus $D(E,U,V^{\mathrm{SC}}) = D^{\mathrm{SA}}(E,U,q^\star)$. Then, under the assumptions of Theorem 2.2, Theorem 2.1 gives that for any $\alpha < 1/D^{\mathrm{SA}}(E,U,q^\star)$ there exist finite cooling schedules of the form (2.6) such that the convergence measure (2.5) satisfies $\mathcal{M}(N) \leqslant N^{-\alpha}$ for N large enough. In other words, piecewise-constant exponential cooling makes it possible for SC to have a convergence speed exponent arbitrarily close to the optimal exponent of the SA algorithm obtained at the limit $N = +\infty$. We conclude this section with some remarks about the conditions for this to occur.

Remark 2. Conditions (iii) and (iv) in Theorem 2.2 are basic assumptions on the communication mechanism in SA theory. They guarantee that any state can be reached from any other state in finitely many steps and that any path in the energy landscape can be followed in the reverse direction.

Remark 3. If condition (v) is not met, it is always possible to replace the family $(q_\beta)_\beta$ with the family $(q'_\beta)_\beta$ defined by

$$q'_\beta(x,y) = \begin{cases} \dfrac{(1-\varepsilon)q_\beta(x,y)}{1-q_\beta(x,x)} & \text{if } y \neq x, \\ \varepsilon & \text{if } y = x, \end{cases}$$

where $\varepsilon \in (0,1)$, so that the algorithm can rest in any state.

Remark 4. Condition (vi) can be rephrased as follows: for any $(x,y) \notin \mathrm{supp}(q^\star)$, $q_\beta(x,y)$ goes to zero faster than any positive power of $\mathrm{e}^{-\beta}$ as $\beta \to +\infty$. A simple sufficient condition for this to hold is that $\mathrm{supp}(q_\beta) = \mathrm{supp}(q^\star)$ for β large enough.

Remark 5. Conditions (i)–(vi) are not sufficient for (2.7) to hold and hence for SC to be "log-optimal" in the sense of (2.8). Using a proof similar to that of Theorem 1 in [51], we

can show that (2.7) is satisfied if (i) is replaced by (2.14) and if (v) and (vi) are replaced by conditions (v′) and (vi′) below.

(v′) For all $\beta \in \mathbb{R}_+$, $\{(x,x);\, x \in E\} \subset \operatorname{supp}(q_\beta) \subset \operatorname{supp}(q^\star)$.

(vi′) For all $(x,y) \in E^2$, the maps $\beta \mapsto q_\beta(x,y)$ and $\beta \mapsto U_\beta(x)$ are continuous.

3. Application to piecewise-constant signal reconstruction. We now consider the problem of minimizing an energy function U of the general form given by (1.2), (1.3), and (1.6) for the purpose of piecewise-constant signal reconstruction. (In the two- or three-dimensional cases, think of x as a lexicographically ordered vector representing an image or a volume.) In particular, the linear operators $\mathcal{D}_i$ in (1.3) are either first-order differences or discrete approximations to the gradient.

Since the theory presented in section 2 assumes a finite state space, we take

$$\Lambda \;=\; \{\, c_1 j + c_2 \,;\; j \in [\![0, L-1]\!]\} \tag{3.1}$$

with $L \in \mathbb{N} \setminus \{0,1\}$ and $(c_1, c_2) \in \mathbb{R}_+^* \times \mathbb{R}$. If Λ^K is a proper subset of the original domain E of U, then minimizing the restriction of U to Λ^K amounts to searching over Λ^K for the best possible approximation of some element of $E_{\min}$ with a level of accuracy determined by the step-size parameter c_1. That being said, everything is about smart design of families $(U_\beta)_{\beta\in\mathbb{R}_+}$ and $(q_\beta)_{\beta\in\mathbb{R}_+}$ satisfying the assumptions of Theorem 2.2.

3.1. Choice of the continuation sequence $(U_\beta)_\beta$. Since the difficulty in minimizing U is due to our choice of ϕ in (1.6), it is natural to focus on SC algorithms in which the continuation sequence $(U_\beta)_\beta$ is obtained by replacing the 0-1 function ϕ in (1.2) with some function parameterized by β; that is,

$$U_\beta(x) \;=\; \mathcal{J}(x) + \lambda \sum_{i \in [\![1,M]\!]} \phi_\beta\big(\|\mathcal{D}_i(x)\|\big),$$

where $(\phi_\beta : \mathbb{R}_+ \to \mathbb{R}_+)_\beta$ is a family of increasing functions such that $\lim_{\beta\to+\infty} \phi_\beta(t) = \phi(t)$ for all t. It is then readily seen that $\lim_{\beta\to+\infty} U_\beta(x) = U(x)$ for all x.

It makes sense that the difficulty in minimizing U_β should be an increasing function of β, although this is not a necessary condition for SC to converge. Assuming that ϕ_β is twice differentiable on $\mathbb{R}_+^*$ for any β, this amounts to saying that the *maximum concavity* of ϕ_β, $\rho(\beta) = \max\{0, -\inf_{t>0} \phi_\beta''(t)\}$, should increase with β. There are various ways to construct such a family $(\phi_\beta)_\beta$ [44], but contrary to GNC, $\rho(\beta)$ does not need to go to zero as $\beta \to 0$. It is also desirable that ϕ_β be not too far from ϕ so that the relaxed energy U_β is reasonably close to the target energy U. Based on these comments, we suggest taking

$$\phi_\beta(t) \;=\; 1 - \frac{1}{1 + (t/\Delta(\beta))^\kappa}\,, \tag{3.2}$$

where $\kappa \in (1, +\infty)$ is fixed and $\Delta : \mathbb{R}_+ \to \mathbb{R}_+$ decreases monotonically to zero. Further motivation for this choice of ϕ_β comes from the fact that it satisfies the assumptions of Theorem 3.1 in [46]; that is, there exist two functions $\tau, \mathcal{T} : \mathbb{R}_+ \to \mathbb{R}_+$ such that the following hold:

1. $0 < \tau < \mathcal{T}$,

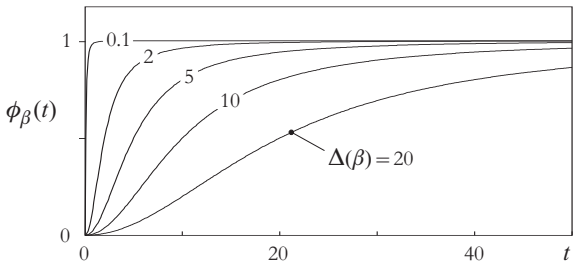

Figure 1. *Some functions in the family $(\phi_\beta)_\beta$ defined by (3.2) with $\kappa = 2$.*

2. $\phi''_\beta \geqslant 0$ on $[0, \tau(\beta)]$ and $\phi''_\beta \leqslant 0$ on $[\tau(\beta), +\infty]$,
3. ϕ''_β is decreasing on $(\tau(\beta), \mathcal{T}(\beta))$ and increasing on $(\mathcal{T}(\beta), +\infty)$.

These functions are given by

$$\tau(\beta) = \Delta(\beta)\left(\frac{\kappa-1}{\kappa+1}\right)^{1/\kappa} \quad \text{and} \quad \mathcal{T}(\beta) = \Delta(\beta)\left(\frac{\kappa-1}{\kappa+2}\left(\frac{\sqrt{3}\kappa}{\sqrt{\kappa^2-1}}+2\right)\right)^{1/\kappa}.$$

If $\mathcal{J}$ is of the form (1.4) and if the $\mathcal{D}_i$'s are difference operators, then there exists $\beta_0 \in \mathbb{R}_+$ and there exist two functions $\theta_0, \theta_1 : \mathbb{R}_+ \to \mathbb{R}_+$ such that, for any $\beta \geqslant \beta_0$, $\tau(\beta) < \theta_0(\beta) < \mathcal{T}(\beta) < \theta_1(\beta)$ and every local minimizer z_β of U_β satisfies either $|\mathcal{D}_i(z_\beta)| < \theta_0(\beta)$ or $|\mathcal{D}_i(z_\beta)| > \theta_1(\beta)$ for any $i \in [\![1, M]\!]$. In other words, $\mathcal{T}(\beta)$ is the edge-detection threshold associated with ϕ_β. Therefore, since $\lim_{\beta\to+\infty} \mathcal{T}(\beta) = 0$, smooth regions in the minimizers of U_β gradually turn into constant regions as $\beta \to +\infty$, while virtually any discontinuity in the true signal can eventually be recovered.

In our experiments in section 4, we use $\kappa = 2$ so that the edge-detection threshold $\mathcal{T}(\beta)$ is equal to the scale parameter $\Delta(\beta)$. Examples of the corresponding functions ϕ_β are shown in Figure 1; they have reversed Lorentzian shape and maximum concavity $\rho(\beta) = \Delta^2(\beta)/2$. Given $\Delta_{\text{sup}} > 0$, we take the function Δ to be a smoothed version of the map $\beta \mapsto \Delta_{\text{sup}} f(\beta)$ with

$$f(\beta) = \begin{cases} 1 & \text{if } \beta < \beta_{\min}, \\ \frac{\beta_{\max}-\beta}{\beta_{\max}-\beta_{\min}} & \text{if } \beta \in [\beta_{\min}, \beta_{\max}], \\ 0 & \text{if } \beta > \beta_{\max}, \end{cases} \tag{3.3}$$

where $\beta_{\min}$ and $\beta_{\max}$ are the minimum and maximum inverse temperature values of the cooling schedule. The smoothing is performed by convolving f in the time domain with a Gaussian function; that is, denoting the convolution operator by $*$,

$$\Delta = \Delta_{\text{sup}}((f \circ \beta^\dagger) * g_\sigma) \circ (\beta^\dagger)^{-1}$$

$$\text{with} \quad g_\sigma(t) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-t}{2\sigma^2}\right) \quad \text{and} \quad \beta^\dagger(t) = \beta_{\min}\left(\frac{\beta_{\max}}{\beta_{\min}}\right)^{\frac{t-1}{N-1}}. \tag{3.4}$$

The problem of choosing an appropriate value for Δ_{sup} is greatly facilitated if we have an estimate ϑ of the maximum discontinuity magnitude in the true signal (as is usually the case

in practice). A good rule of thumb is to take $\Delta_{\sup} \geqslant 2\vartheta$ so that for any relevant minimum z of U the set $\{\mathcal{D}_i(z)\,;\ i \in [\![1, M]\!]\}$ is in the convex region of ϕ_β at the beginning of the SC process.

3.2. Design of the communication sequence $(q_\beta)_\beta$. Since we are looking for piecewise-constant solutions, it is worthwhile to restrict the domain Λ^K of U to a *locally bounded space*, that is, a set Ω_ζ which consists of the elements $x = (x_1, \dots, x_K)$ of Λ^K such that

$$\forall k \in [\![1, K]\!], \qquad \min_{l \in \mathcal{N}(k)} x_l - c_1\zeta \ \leqslant\ x_k \ \leqslant \max_{l \in \mathcal{N}(k)} x_l + c_1\zeta\,, \tag{3.5}$$

where $\zeta \in [\![1, L-1]\!]$ and $\mathcal{N}$ is a predefined neighborhood system on $[\![1, K]\!]$ (e.g., the $(2r)$ or $(3^r - 1)$ nearest-neighbor systems in the r-dimensional case). Informally, Ω_ζ is the set of configurations in which each component is in the range delimited by its neighbors with some "slack" characterized by ζ. Clearly, this definition does not preclude the presence of discontinuities, and Ω_ζ contains all piecewise-constant configurations in Λ^K whatever the value of ζ. From now on, we take the state space to be Ω_ζ.

The construction of a symmetric and irreducible Markov matrix q on Ω_ζ is detailed in [64]; basically,

$$q(x, y) \ = \begin{cases} \big(K\,|\Omega_\zeta^k(x)|\big)^{-1} & \text{if } y \in \Omega_\zeta^k(x), \\ 0 & \text{otherwise,} \end{cases}$$

$$\text{where} \qquad \Omega_\zeta^k(x) \ = \ \big\{y \in \Omega_\zeta \,\big|\, \forall l \neq k,\ y_l = x_l\big\}.$$

This communication mechanism yields significant improvement in the performance of SA when the energy function promotes solutions that are reasonably smooth in the sense of (3.5) [64, 53]. In our case, the target energy is more selective in that it encourages piecewise-constant configurations. In particular, each component of a relevant minimum z of U is expected to have at least one neighboring component with the same value, that is,

$$\forall k \in [\![1, K]\!], \quad \exists l \in \mathcal{S}(k), \quad z_l = z_k, \tag{3.6}$$

where $\mathcal{S}$ is the neighborhood system on $[\![1, K]\!]$ defined as follows:

$$l \in \mathcal{S}(k) \iff \exists i \in [\![1, M]\!],\ \mathcal{D}_i(e^{(l)}) \neq 0,\ \text{and}\ \mathcal{D}_i(e^{(k)}) \neq 0,$$

where $e^{(k)}$ denotes the kth vector of the standard basis of $\mathbb{R}^K$. To take advantage of this characteristic, we propose to use a communication sequence of the form

$$q_\beta \ = \ (1 - \Theta(\beta))q + \Theta(\beta)\widetilde{q}\,,$$

where $\Theta : \mathbb{R}_+ \to [0, 1]$ is monotonic increasing and where $\widetilde{q}$ is a Markov matrix designed to favor the formation of configurations satisfying (3.6). The function Θ gives the probability of choosing $\widetilde{q}$ rather than q to generate a new candidate solution. It is assumed to be increasing because the visited states are expected to be nearly piecewise-constant by the end of the SC process. In addition, $\Theta(\beta)$ should be close to zero for small values of β to ensure efficient

exploration of the state space at the beginning of the SC process. We take Θ to be a smoothed version of the map $\beta \mapsto \Theta_{\text{sup}}(1 - f(\beta))$ with $\Theta_{\text{sup}} \in (0,1)$ and where f is given by (3.3); that is,

$$\Theta \;=\; \Theta_{\text{sup}}\big(1 - ((f \circ \beta^{\dagger}) * g_{\sigma}) \circ (\beta^{\dagger})^{-1}\big),$$

where g_{σ} and $\beta^{\dagger}$ are defined in (3.4). Then, since q is irreducible and $q(x,x) > 0$ for all $x \in \Omega_{\zeta}$, the limit kernel

$$q^{\star} \;=\; \lim_{\beta \to +\infty} q_{\beta} \;=\; (1 - \Theta_{\text{sup}})q + \Theta_{\text{sup}}\widetilde{q}$$

satisfies conditions (iii) and (v) in Theorem 2.2, and it remains only to construct $\widetilde{q}$ while guaranteeing that conditions (iv) and (vi) are met.

We propose that $\widetilde{q}$ generates a new candidate solution y from x by replacing the value of x at a randomly selected location k with a neighboring value in the set

$$\Upsilon^{k}(x) = \big\{x_l \,\big|\, l \in \mathcal{S}(k)\big\} \cap \big\{z_k \,\big|\, z \in \Omega_{\zeta}^{k}(x)\big\}\,.$$

More specifically, $y_l = x_l$ for all $l \neq k$, and y_k is drawn from the probability distribution p_x^k on $\Upsilon^k(x)$ defined by

$$p_x^k(\omega) \;=\; \frac{\sum_{l \in \mathcal{S}(k)} \mathbb{1}_{\{x_l = \omega\}} / \xi(k,l)}{\sum_{l \in \mathcal{S}(k)} \mathbb{1}_{\{x_l \,\in\, \Upsilon^k(x)\}} / \xi(k,l)}\,, \tag{3.7}$$

where ξ is a metric on $[\![1, K]\!]$. (In simple terms, $p_x^k(\omega)$ reflects the likelihood that $x_k = \omega$ in the event that x is piecewise-constant.) It may happen that $\Upsilon^k(x) = \emptyset$, in which case we set $y_k = x_k$. Thus,

$$\widetilde{q}(x,y) \;=\; \begin{cases} K^{-1}\, p_x^k(y_k) & \text{if } y_l = x_l \quad \forall l \neq k \text{ and if } y_k \in \Upsilon^k(x), \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that $\text{supp}(\widetilde{q}) \subset \text{supp}(q)$. Consequently, $\text{supp}(q_{\beta}) = \text{supp}(q) = \text{supp}(q^{\star})$ for any β, and hence condition (iv) follows from the symmetry of q and condition (vi) follows from Remark 4. Note that $\widetilde{q}$ is not irreducible, as there is no $\widetilde{q}$-admissible path between any two configurations that do not satisfy (3.6). This is the reason why we require the upper bound Θ_{sup} be smaller than 1.

4. Experiments. The experiments presented here concern the problem of reconstructing a piecewise-constant three-dimensional test object from a few noisy two-dimensional line-integral projections. The test object is depicted in Figure 2; it consists of a sphere S inside a regular octahedron O, the latter being contained in a cube C. The corresponding true configuration $x^{\star}$ is defined over a voxel grid of size $K = 49^3$. Denoting the center of the kth voxel by $\mathsf{c}(k)$, we have $x_k^{\star} = 30$ if $\mathsf{c}(k) \in \mathsf{S}$, $x_k^{\star} = 20$ if $\mathsf{c}(k) \in \mathsf{O} \setminus \mathsf{S}$, $x_k^{\star} = 10$ if $\mathsf{c}(k) \in \mathsf{C} \setminus \mathsf{O}$, and $x_k^{\star} = 0$ if $\mathsf{c}(k) \notin \mathsf{C}$. The data are shown in Figure 3. They consist of six 128×128 simulated cone-beam projections corrupted by white Gaussian noise at 20 dB signal-to-noise ratio (SNR). The associated source positions form the vertices of a regular octahedron, and the decibel level of the SNR is $10 \log_{10}\big(\varsigma^2 / \sigma_{\eta}^2\big)$, where σ_{η}^2 is the variance of the noise and ς^2 is the variance of the exact data $\mathcal{H}(x^{\star})$.

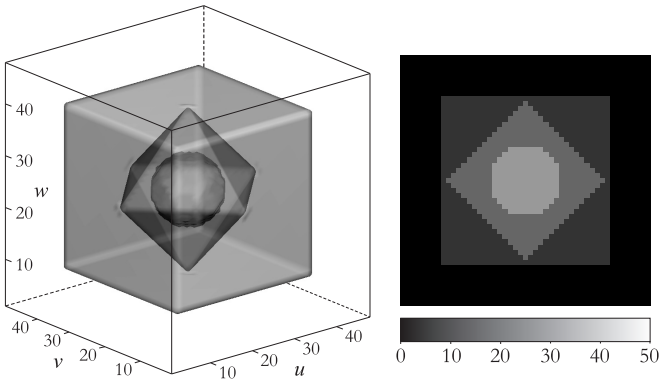

Figure 2. *Three-dimensional test object: isosurface representation and vw cross section at* $u = 24$.

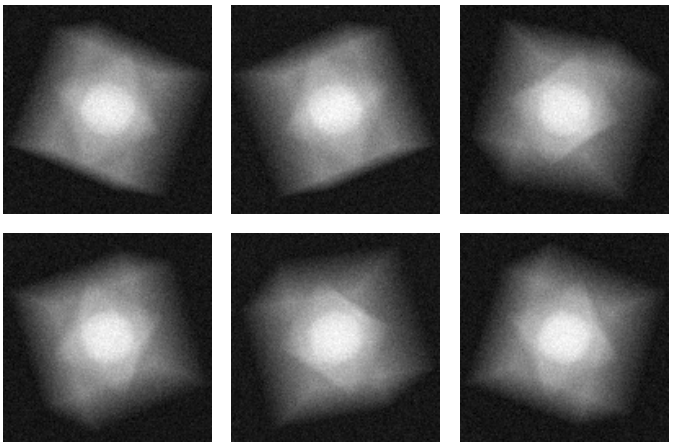

Figure 3. *Simulated line-integral projection data associated with the object shown in Figure 2.*

The solution is defined as any global minimum of the energy

$$U : x \in \Lambda^K \longmapsto \frac{1}{\sigma_\eta^2} \|\mathcal{H}(x) - d\|^2 + \lambda \sum_{\{k,l\} \in \mathcal{C}} \phi\big(|x_k - x_l|\big), \tag{4.1}$$

where the set of admissible voxel values Λ in (3.1) is defined by $L = 50$ and $(c_1, c_2) = (1, 0)$, $\|\cdot\|$ is the ℓ_2-norm, ϕ is the 0-1 function in (1.6), $\mathcal{C}$ is the set of pair-site cliques associated with the 26 nearest-neighbor system, and the value of the smoothing parameter λ is computed as described in [51] ($\lambda = 1.06$). This function is an important special case of the class of energies defined in the introduction. Note, however, that our SC algorithm described in section 3 applies equally well to any energy of the general form

$$U(x) = \mathcal{J}(x) + \lambda \sum_{i \in [\![1,M]\!]} \phi\big(\|\mathcal{D}_i(x)\|\big),$$

where the $\mathcal{D}_i$'s are any linear maps and whether or not the data-likelihood function $\mathcal{J}$ is convex.

4.1. The competing algorithms. We investigate the behavior of six different algorithms—two deterministic continuation algorithms, two SA algorithms, and two SC algorithms—to make a three-way comparison between standard continuation, standard MCMC optimization, and SC.

The families of relaxed energies of the two deterministic continuation (DC) algorithms are obtained by replacing ϕ with the elements of a sequence $(\phi_n)_{n\in\mathbb{N}^*}$ of potential functions converging pointwise to ϕ and starting with ϕ_1 convex on $[0,50]$. (This interval covers the range of admissible voxel values.) The first considered continuation sequence (algorithm DC_1) is of the same form as for the SC approach, and the second continuation sequence (algorithm DC_2) is the one originally proposed by Leclerc [38]. More precisely, the relaxation scheme of DC_1 is defined by

$$\phi_n(t) \;=\; 1 - \frac{1}{1+(t/\Delta_n)^2}\,, \tag{4.2}$$

where $(\Delta_n)_{n\in\mathbb{N}^*}$ is a sequence of positive reals decreasing to zero, and the relaxed potentials of DC_2 are given by

$$\phi_n(t) \;=\; 1 - \exp\bigl(-(t/\Delta_n)^2\bigr). \tag{4.3}$$

Following [38], we take $(\Delta_n)_n$ to be of the form

$$\Delta_n \;=\; \Delta_1 \chi^{n-1},$$

where Δ_1 is chosen large enough to ensure the convexity of ϕ_1 on $[0,50]$ (i.e., $\Delta_1 = 50\sqrt{3}$ for DC_1 and $\Delta_1 = 50\sqrt{2}$ for DC_2) and where the parameter $\chi \in (0,1)$ controls the speed of convergence to the target energy. We set χ to 0.95, and the relaxation process ends with the completion of the first iteration n for which $\phi_n(1) \geqslant 0.99$ (increasing χ or the number of iterations further does not yield any noticeable improvement). Finally, the optimization task that takes place at each iteration is performed by half-quadratic minimization [15].

The stochastic algorithms under consideration are SA with logarithmic cooling (algorithm SA_1), SA with exponential cooling (algorithm SA_2), SC with time-dependent energy and fixed communication (algorithm SC_1), and SC with time-dependent energy and time-dependent communication (algorithm SC_2), that is,

$$\begin{aligned} \mathrm{SA}_1 &= (\Omega_\zeta, q, U, (B\ln(n+1))_{n\in[\![1,N]\!]}),\\ \mathrm{SA}_2 &= (\Omega_\zeta, q, U, (\beta_n)),\\ \mathrm{SC}_1 &= (\Omega_\zeta, q, (U_\beta), (\beta_n)),\\ \text{and}\qquad \mathrm{SC}_2 &= (\Omega_\zeta, (q_\beta), (U_\beta), (\beta_n)), \end{aligned}$$

where the state space Ω_ζ, the communication mechanism q, and the sequences $(U_\beta)_\beta$ and $(q_\beta)_\beta$ are as specified in section 3. The locally bounded space Ω_ζ is defined by the 6-nearest neighbor system and $\zeta = 2$, the parameters Δ_{sup} and Θ_{sup} are set to 20.0 and 0.95, respectively, and $\xi(k,l)$ in (3.7) is the Euclidean distance between the centers of the kth and lth voxels. Algorithms SA_2, SC_1, and SC_2 use piecewise-constant exponential cooling of the form (2.6)

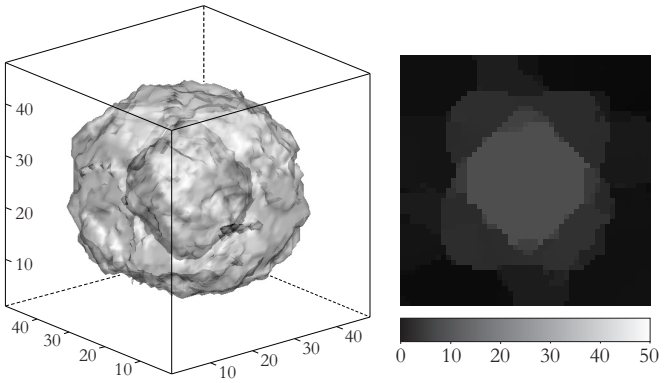

Figure 4. *Reconstruction using DC with the Lorentzian-shape relaxation scheme* (4.2) *(algorithm* DC_1*):* $RMSE = 3.5756$, $U = 1.5082 \cdot 10^6$.

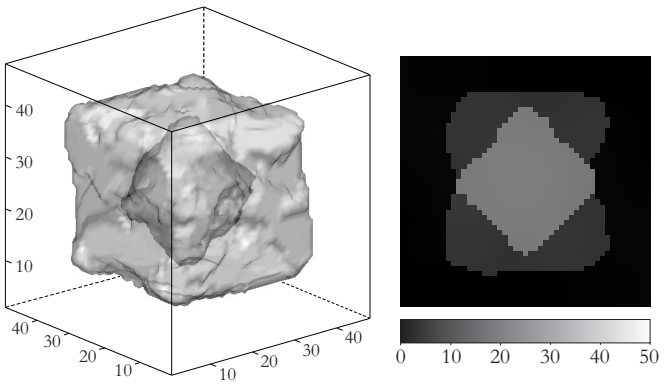

Figure 5. *Reconstruction using DC with the Leclerc relaxation scheme* (4.3) *(algorithm* DC_2*):* $RMSE = 2.3190$, $U = 3.1005 \cdot 10^5$.

with $(N, \nu) = (10^4 K, 500)$ and with initial and final temperature values selected by means of the procedures proposed in [53]. The horizon N of SA_1 is also set to $10^4 K$, and the positive constant B of the logarithmic schedule is chosen so that SA_1 ends at the same temperature as SA_2. All four algorithms start from random configurations.

4.2. Results. The reconstructions obtained by standard continuation are shown in Figures 4 and 5; both algorithms completely miss the sphere and fail to recover the cube properly. The relaxation method of Leclerc (algorithm DC_2) performs much better than the Lorentzian-shape relaxation scheme (algorithm DC_1): the root-mean-square errors (RMSE) are respectively 3.5756 and 2.3190 for DC_1 and DC_2, and the energy of the reconstructed object is substantially larger for DC_1 $(1.5082 \cdot 10^6)$ than for DC_2 $(3.1005 \cdot 10^5)$. The computation time was 3 hours and 4 minutes for DC_1 and 2 hours and 18 minutes for DC_2 on a Quad-Core Xeon 2.66 GHz (L5430) machine.

The results associated with the stochastic algorithms are summarized in Table 1. Each algorithm was run 30 times in order to assess the variability of the estimates inherent to the stochastic approach. In any case, the computation time for a single run did not exceed 45

Table 1

Statistics over 30 runs of SA with logarithmic cooling (algorithm SA_1), SA with exponential cooling (algorithm SA_2), SC with time-dependent energy and fixed communication (algorithm SC_1), and SC with time-dependent energy and time-dependent communication (algorithm SC_2). Given are the minimum, maximum, mean, and standard deviation of the RMSE and of the final energy.

Alg.	RMSE				Final energy U			
	Min	Max	Mean	Stdv	Min	Max	Mean	Stdv
SA_1	6.780	7.219	6.983	1.189	$3.3729 \cdot 10^5$	$3.5294 \cdot 10^5$	$3.4418 \cdot 10^5$	4155.4
SA_2	8.700	9.020	8.853	0.0764	$3.3414 \cdot 10^5$	$3.4002 \cdot 10^5$	$3.3693 \cdot 10^5$	1417.4
SC_1	1.432	1.534	1.472	0.0249	$2.5668 \cdot 10^5$	$2.6555 \cdot 10^5$	$2.6191 \cdot 10^5$	1828.5
SC_2	0.382	0.415	0.400	0.0075	$1.8186 \cdot 10^5$	$1.8189 \cdot 10^5$	$1.8187 \cdot 10^5$	8.8

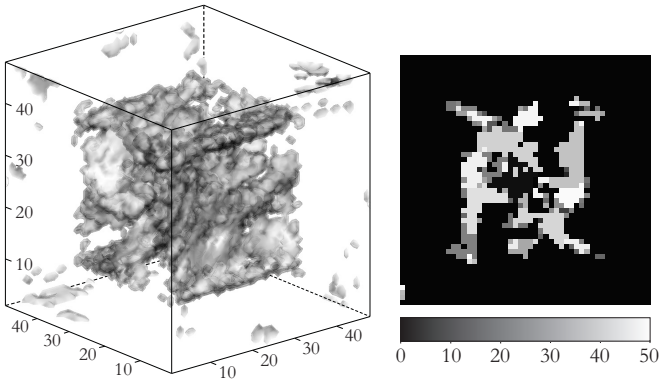

Figure 6. *Reconstruction by SA. Best solution in terms of energy: $U = 3.3414 \cdot 10^5$.*

minutes. We observe that DC with Leclerc's relaxation greatly outperforms SA regardless of the cooling schedule. The best reconstruction produced by SA is actually a very poor local minimum; it is shown in Figure 6. A natural question that arises is whether increasing the length N of the annealing chain can yield significant improvements. The answer is negative. Indeed, using SA_2 with 400 times more iterations (i.e., $N = 4 \cdot 10^6 K$ and about 12 days of computation), we obtained a solution with an RMSE of 9.258 (which is even greater than the RMSE of the estimates produced by SA in $10^4 K$ iterations!) and an energy of $3.0509 \cdot 10^5$ (which is slightly smaller than the energy obtained by DC, but 15% greater than the energy of the worst solution found by SC in $10^4 K$ iterations). This shows that the energy landscape (Ω_ζ, U, q) is difficult in the sense that it contains poor local minima with deep basins of attraction. The situation is in fact even worse: there exist elements of the state space that are far away from the true configuration $x^\star$ but whose energy is very close to the energy of $x^\star$ ($U(x^\star) = 1.8240 \cdot 10^5$). An example is given in Figure 7. This minimum was obtained by SC with fixed energy and time-dependent communication; its energy is only 0.24% above $U(x^\star)$.

On the other hand, SC with time-dependent energy ends up in relevant basins of attraction and substantially outperforms DC (and hence SA). The worst estimate produced by SC_1 is shown in Figure 8. Its energy is 14% smaller than the energy level reached by DC_2, and it is

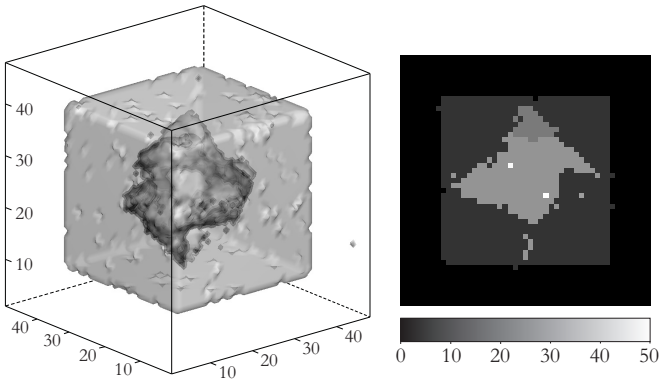

Figure 7. *Undesirable minimum with energy close to $U(x^\star)$: RMSE $= 2.061$, $U = 1.8283 \cdot 10^5$.*

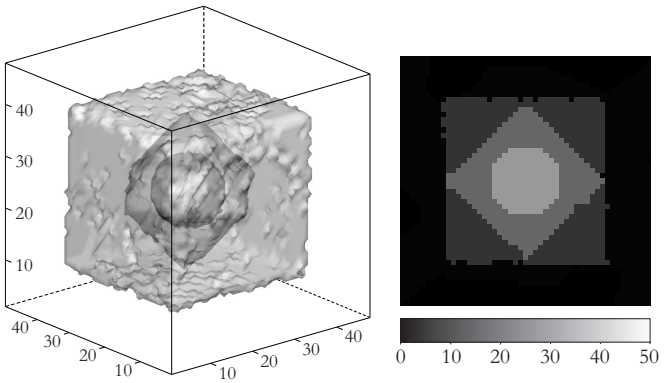

Figure 8. *Reconstruction using SC with time-dependent energy and fixed communication (algorithm SC_1). Worst solution in terms of both RMSE and energy: RMSE $= 1.534$, $U = 2.6555 \cdot 10^5$.*

closer to $x^\star$ than are the output of DC_2 and the undesirable minimum in Figure 7. Further significant improvements are obtained by allowing the communication mechanism to be time-dependent. The results obtained by algorithm SC_2 are striking: the maximum RMSE is only 0.83% of the voxel value range, and the standard deviation of the final energy is negligible (about 0.005% of the mean final energy). Moreover, in all runs, the energy of the computed solution turns out to be slightly smaller than $U(x^\star)$. The worst solution computed by SC_2 is shown in Figure 9. It is almost identical to the true configuration, and its energy is 26% smaller than the energy of the best estimate produced by SC_1.

5. Conclusion. We introduced a new class of hybrid algorithms, namely stochastic continuation (SC), which provides great freedom in the design of annealing-type algorithms. SC is interesting in several respects. First, SC inherits the convergence properties of generalized simulated annealing (SA) under weak assumptions and is therefore more theoretically grounded than deterministic continuation (DC). Second, well-designed SC algorithms can substantially outperform SA without requiring additional computational efforts. Third, the scope of SC is not limited to specific classes of energies.

Our experimental results in the context of signal reconstruction showed that SC is a

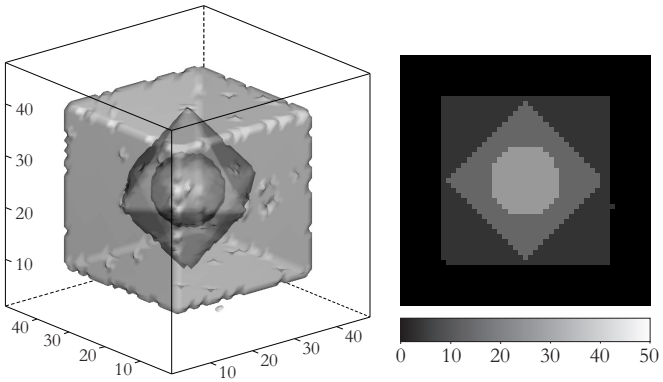

Figure 9. *Reconstruction using SC with time-dependent energy and time-dependent communication (algorithm* SC_2*). Worst solution in terms of both RMSE and energy: RMSE* $= 0.415$*,* $U = 1.8189 \cdot 10^5$*.*

valuable alternative when both DC and SA fail. More generally, the flexibility of SC makes it potentially attractive for a wide range of difficult optimization problems in the computer vision and signal processing fields.

Appendix A. A fundamental limitation of standard graph cuts. In [37], Kolmogorov and Zabih give a necessary condition for a *pseudo-Boolean function* (i.e., a mapping from $\mathbb{B}^K$ to $\mathbb{R}$, where $\mathbb{B} := \{0,1\}$) to be minimized by standard graph cuts. Here, we show that this condition is not satisfied by simple energy functions arising in MAP-MRF reconstruction. We consider a particular case of the class of functions defined by (1.2) and (1.3), namely,

$$U : x \in \mathbb{B}^K \longmapsto \|\mathcal{H}(x) - d\|^2 + \lambda \sum_{\{k,l\} \in \mathcal{C}} |x_k - x_l|\,, \tag{A.1}$$

where $\|\cdot\|$ is the ℓ_2-norm, $\mathcal{H} : \mathbb{R}^K \to \mathbb{R}^{K'}$ is a linear map, $d \in \mathbb{R}^{K'}$, $\lambda \in \mathbb{R}_+^*$, and $\mathcal{C} = \{\{k,l\} \,|\, l \in \mathcal{S}(k)\}$ is the set of pair-site cliques associated with a neighborhood system $\mathcal{S}$ on $[\![1,K]\!]$.

Given $\alpha \in \mathbb{B}^{K-2}$, $(a,b) \in \mathbb{B}^2$, and $(i,j) \in [\![1,K]\!]^2$ such that $i \neq j$, we denote by $x_{i,j}^{\alpha}(a,b)$ the element $(x_1,\ldots,x_K) \in \mathbb{B}^K$ defined by

$$\begin{cases} x_i = a, \;\; x_j = b, \\ \forall k \in [\![1,K]\!] \setminus \{i,j\}, \quad x_k = \alpha_{r_{i,j}(k)}, \end{cases}$$

with $r_{i,j}(k) = k - (\mathbb{1}_{\{i<k\}} + \mathbb{1}_{\{j<k\}})$. Let $U_{i,j}^{\alpha} : \mathbb{B}^2 \to \mathbb{R}$ be defined by $U_{i,j}^{\alpha}(a,b) = U(x_{i,j}^{\alpha}(a,b))$. Then U is said to be *submodular* if

$$U_{i,j}^{\alpha}(0,0) + U_{i,j}^{\alpha}(1,1) \leqslant U_{i,j}^{\alpha}(0,1) + U_{i,j}^{\alpha}(1,0)$$

for any $\alpha \in \mathbb{B}^{K-2}$ and for any $(i,j) \in [\![1,K]\!]^2$ such that $i \neq j$. It is shown in [37] that a pseudo-Boolean function cannot be minimized by graph cuts if it is not submodular. (Actually, the limitation of pseudo-Boolean optimization caused by nonsubmodularity was known prior to [37]: see, e.g., [5, section 6.1].) Let $\langle \cdot,\cdot \rangle$ be the Euclidean inner product, and let $e^{(i)}$ denote the ith vector of the standard basis of $\mathbb{R}^K$. Proposition A.1 below gives a necessary and sufficient

condition for U to be submodular, and hence a necessary condition for U to be minimized by graph cuts. In particular, U is not submodular if there exists $(i,j) \in [\![1,K]\!]^2$ such that $i \neq j$, $j \notin \mathcal{S}(i)$, and $\langle \mathcal{H}(e^{(i)}), \mathcal{H}(e^{(j)}) \rangle > 0$. Since most of the time $\mathcal{S}$ is a nearest-neighbor system with small support, it follows that graph cuts cannot deal with the fundamental cases where $\mathcal{H}$ is a line-integral projection operator and where $\mathcal{H}$ is a convolution operator whose kernel has support larger than $\mathcal{S}$.

Proposition A.1. *The function U defined in (A.1) is submodular if and only if, for any $(i,j) \in [\![1,K]\!]^2$ such that $i \neq j$,*

$$\langle \mathcal{H}(e^{(i)}), \mathcal{H}(e^{(j)}) \rangle \leqslant \begin{cases} \lambda & \text{if } j \in \mathcal{S}(i), \\ 0 & \text{otherwise.} \end{cases}$$

Proof. Let $\mathcal{J}$ and Φ be the functions from $\mathbb{B}^K$ to $\mathbb{R}$ defined by

$$\mathcal{J}(x) = \|\mathcal{H}(x) - d\|^2 \qquad \text{and} \qquad \Phi(x) = \sum_{\{k,l\} \in \mathcal{C}} |x_k - x_l| .$$

Let $\alpha \in \mathbb{B}^{K-2}$, $(a,b) \in \mathbb{B}^2$, and $(i,j) \in [\![1,K]\!]^2$ such that $i \neq j$. We can write

$$x^\alpha_{i,j}(a,b) = x_{i,j}(a,b) + x^\alpha_{i,j}$$
$$\text{with} \quad x_{i,j}(a,b) = ae^{(i)} + be^{(j)} \quad \text{and} \quad x^\alpha_{i,j} = \sum_{k \in [\![1,K]\!] \setminus \{i,j\}} \alpha_{r_{i,j}(k)} e^{(k)},$$

and it is not difficult to check that

$$\mathcal{J}(x^\alpha_{i,j}(a,b)) = \|\mathcal{H}(x_{i,j}(a,b))\|^2 + 2\,\langle\, \mathcal{H}(x_{i,j}(a,b)), \mathcal{H}(x^\alpha_{i,j}) - d \,\rangle + c_{\mathcal{J}}, \tag{A.2}$$

where $c_{\mathcal{J}}$ is a constant independent of a and b. For any $x \in \mathbb{B}^K$,

$$\Phi(x) = \sum_{\{k,l\} \in \mathcal{C}_{i,j}} |x_k - x_l| + \Psi\Big(x_{r^{-1}_{i,j}(1)}, \ldots, x_{r^{-1}_{i,j}(K-2)}\Big),$$

where $\mathcal{C}_{i,j} = \{\{k,l\} \,|\, l \in \mathcal{S}(k) \text{ and } \{k,l\} \cap \{i,j\} \neq \emptyset\}$ and where

$$\Psi\Big(x_{r^{-1}_{i,j}(1)}, \ldots, x_{r^{-1}_{i,j}(K-2)}\Big) = \sum_{\{k,l\} \in \mathcal{C} \setminus \mathcal{C}_{i,j}} |x_k - x_l| .$$

The set $\mathcal{C}_{i,j}$ can be written as a disjoint union:

$$\mathcal{C}_{i,j} = \big\{\{i,k\} \,\big|\, k \in \mathcal{S}(i) \setminus \{j\}\big\} \cup \big\{\{j,k\} \,\big|\, k \in \mathcal{S}(j) \setminus \{i\}\big\} \cup \mathcal{B}_{i,j}$$
$$\text{with} \qquad \mathcal{B}_{i,j} = \begin{cases} \{\{i,j\}\} & \text{if } j \in \mathcal{S}(i), \\ \emptyset & \text{otherwise.} \end{cases}$$

Consequently, since the function Ψ is independent of x_i and x_j,

$$\Phi(x^\alpha_{i,j}(a,b)) = \sum_{k \in \mathcal{S}(i) \setminus \{j\}} |a - \alpha_{r_{i,j}(k)}| + \sum_{k \in \mathcal{S}(j) \setminus \{i\}} |b - \alpha_{r_{i,j}(k)}| + \mathbb{1}_{\{j \in \mathcal{S}(i)\}} |a - b| + c_\Phi, \tag{A.3}$$

where c_Φ is a constant independent of a and b.

Using (A.2) and (A.3), we can compute $U^\alpha_{i,j}(a,b) = (\mathcal{J} + \lambda\Phi)(x^\alpha_{i,j}(a,b))$ for $(a,b) = (0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$. Substituting these results into

$$U^\alpha_{i,j}(0,0) + U^\alpha_{i,j}(1,1) - U^\alpha_{i,j}(0,1) - U^\alpha_{i,j}(1,0) \;=:\; \Delta^\alpha_{i,j}(U),$$

we obtain, after simplification,

$$\Delta^\alpha_{i,j}(U) \;=\; 2\Big(\langle \mathcal{H}(e^{(i)}), \mathcal{H}(e^{(j)})\rangle - \lambda \mathbb{1}_{\{j\in\mathcal{S}(i)\}}\Big)$$

(note that $\Delta^\alpha_{i,j}(U)$ is independent of α). By definition, U is submodular if and only if $\Delta^\alpha_{i,j}(U) \leqslant 0$ for any $\alpha \in \mathbb{B}^{K-2}$ and for any $(i,j) \in [\![1,K]\!]^2$ such that $i \neq j$. Hence the proposition follows. ∎

Appendix B. On the performance of QPBO for binary signal reconstruction. Consider the simple class of pseudo-Boolean functions defined in (A.1); these are of the form (1.1) with pairwise interactions

$$\phi_{\{k,l\}} : (a,b) \in \mathbb{B}^2 \longmapsto 2\langle \mathcal{H}(e^{(k)}), \mathcal{H}(e^{(l)})\rangle\, ab \;+\; \mathbb{1}_{\{\{k,l\}\in\mathcal{C}\}}\lambda|a-b|\,,$$

where $\{k,l\}$ is any 2-subset of $[\![1,K]\!]$, $\langle\cdot,\cdot\rangle$ is the Euclidean inner product, and $e^{(k)}$ denotes the kth vector of the standard basis of $\mathbb{R}^K$. Assuming that $\mathcal{H}$ is nonnegative, as is typically the case in vision, the set $\mathcal{R}$ of submodular terms is given by

$$\mathcal{R} \;=\; \big\{\{k,l\} \in \mathcal{C} \;\big|\; \langle \mathcal{H}(e^{(k)}), \mathcal{H}(e^{(l)})\rangle \;\leqslant \lambda\big\},$$

and the set $\overline{\mathcal{R}}$ of nonsubmodular terms consists of the pairs $\{k,l\}$ such that $k \neq l$ and

$$\begin{cases} \langle \mathcal{H}(e^{(k)}), \mathcal{H}(e^{(l)})\rangle \;>\; \lambda & \text{if } \{k,l\} \in \mathcal{C}, \\ \langle \mathcal{H}(e^{(k)}), \mathcal{H}(e^{(l)})\rangle \;>\; 0 & \text{if } \{k,l\} \notin \mathcal{C}. \end{cases}$$

In signal reconstruction applications of QPBO methods, it is general practice to choose λ large enough to have a maximum number of submodular terms, i.e., $\lambda \geqslant \max_{\{k,l\}\in\mathcal{C}}\langle \mathcal{H}(e^{(k)}), \mathcal{H}(e^{(l)})\rangle$ so that $\mathcal{R} = \mathcal{C}$. In this case, $\overline{\mathcal{R}} = \mathcal{C}_\mathcal{H}$ with

$$\mathcal{C}_\mathcal{H} \;=\; \big\{\{k,l\} \;\big|\; k \neq l \text{ and } \langle \mathcal{H}(e^{(k)}), \mathcal{H}(e^{(l)})\rangle > 0\big\},$$

and the nonsubmodularity ratio $|\overline{\mathcal{R}}|/|\mathcal{R}|$ satisfies

$$\frac{|\mathcal{C}_\mathcal{H}|}{|\mathcal{C}|} - 1 \;\leqslant\; \frac{|\overline{\mathcal{R}}|}{|\mathcal{R}|} \;\leqslant\; \frac{|\mathcal{C}_\mathcal{H}|}{|\mathcal{C}|}\,.$$

The behavior of QPBO is thus closely linked to the *connectivity ratio* $\varrho = |\mathcal{C}_\mathcal{H}|/|\mathcal{C}|$, that is, the ratio of the number of pair-site cliques in the data-likelihood to the number in the prior. The approach shows good performance when the connectivity ratio is small (i.e., $\varrho \lessapprox 1$). A nice example of its application to parallel magnetic resonance imaging can be found in [50] ($\varrho = (m-1)/8$ for an m-fold acceleration and an 8-nearest neighbor system in the prior).

However, the performance of QPBO decreases rather rapidly as ϱ increases beyond 1. To fix ideas, the binary image restoration experiments reported in [56] show that, for $\varrho = 3$, SA performs similarly to QPBO both in terms of quality of the results and in terms of computation time (in the case of image deconvolution, $\varrho = m(2m+1)$ for a $(2m+1)\times(2m+1)$ point spread function and a 4-nearest neighbor system in the prior). On the other hand, for $\varrho = 10$, QPBO does not find a global minimum—SA does—and is, moreover, substantially slower than SA (more than 20 times slower, actually). What emerges from these observations is that, when applied to signal reconstruction problems with large connectivity ratios (e.g., $\varrho \gtrapprox 10$), QPBO not only performs poorly but also is significantly outperformed by SA. This is especially the case for reconstruction from line-integral projections, where ϱ increases approximately linearly with the number of projections and with the square root of the number of pixels (two-dimensional case) or the cubic root of the number of voxels (three-dimensional case). For instance, $\varrho \approx 33$ for the experimental setup described in section 4, even though there are only six projections!

Appendix C. The particular case of simulated annealing. Here, we show that the family $(Q_\beta^{\mathrm{SA}})_\beta$ defined by (2.2) has rare transitions with rate function V^{SA} given in (2.3). We proceed by cases.

Case 1. If $y \neq x$, then

$$V^{\mathrm{SA}}(x,y) = \begin{cases} (U(y)-U(x))^+ & \text{if } q(x,y) > 0, \\ +\infty & \text{if } q(x,y) = 0, \end{cases}$$

and since

$$\lim_{\beta\to+\infty} -\beta^{-1}\ln q(x,y) = \begin{cases} 0 & \text{if } q(x,y) > 0, \\ +\infty & \text{if } q(x,y) = 0, \end{cases}$$

then

$$\lim_{\beta\to+\infty} -\beta^{-1}\ln Q_\beta^{\mathrm{SA}}(x,y) = \lim_{\beta\to+\infty} -\beta^{-1}\ln q(x,y) + (U(y)-U(x))^+ = V^{\mathrm{SA}}(x,y).$$

Case 2. If $y = x$ and $q(x,x) > 0$, then for any $\beta > 0$

$$1 \geqslant Q_\beta^{\mathrm{SA}}(x,x) \geqslant 1 - \sum_{z\neq x} q(x,z) = q(x,x) > 0.$$

It follows that $V^{\mathrm{SA}}(x,x) = 0$ and that

$$0 \leqslant -\beta^{-1}\ln Q_\beta^{\mathrm{SA}}(x,x) \leqslant -\beta^{-1}\ln q(x,x) \underset{\beta\to+\infty}{\longrightarrow} 0.$$

Hence $\lim_{\beta\to+\infty} -\beta^{-1}\ln Q_\beta^{\mathrm{SA}}(x,x) = V^{\mathrm{SA}}(x,x)$.

Case 3. If $y = x$ and $q(x,x) = 0$, we have to distinguish two subcases depending on whether or not x is a local maximum of the energy landscape (E,U,q).

$\diamond$ If x is a local maximum of (E,U,q), that is, if for all $z \in E$, $q(x,z) > 0 \Longrightarrow U(z) \leqslant U(x)$, then for any $\beta > 0$

$$Q_\beta^{\mathrm{SA}}(x,x) = 1 - \sum_{z\neq x} q(x,z) = q(x,x) = 0,$$

and thus $\lim_{\beta\to+\infty} -\beta^{-1}\ln Q_\beta^{\mathrm{SA}}(x,x) = +\infty = V^{\mathrm{SA}}(x,x)$.

$\diamond$ If x is not a local maximum of (E,U,q), then we can pick $z_0 \in E$ such that $q(x,z_0)>0$ and $U(z_0)>U(x)$. For any $\beta>0$, we have

$$\begin{aligned} 1 \geqslant Q_\beta^{\mathrm{SA}}(x,x) &\geqslant 1 - \sum_{z\in E\setminus\{x,z_0\}} q(x,z) - q(x,z_0)\exp\bigl(-\beta(U(z_0)-U(x))\bigr) \\ &= q(x,z_0)\bigl(1-\exp\bigl(-\beta(U(z_0)-U(x))\bigr)\bigr). \end{aligned}$$

It follows that $Q_\beta^{\mathrm{SA}}(x,x)>0$ for all $\beta>0$ (hence $V^{\mathrm{SA}}(x,x)=0$) and that

$$\lim_{\beta\to+\infty} -\beta^{-1}\ln Q_\beta^{\mathrm{SA}}(x,x) = 0.$$

Appendix D. Proof of Theorem 2.2. The function V^{SC} is irreducible, as $q^\star$ is irreducible and $V^{\mathrm{SC}}(x,y)<+\infty$ if and only if $q^\star(x,y)>0$. Furthermore, since $q^\star$ has symmetric support, $V^{\mathrm{SC}}(x,y)<+\infty$ if and only if $V^{\mathrm{SC}}(y,x)<+\infty$, and thus if $V^{\mathrm{SC}}(x,y)<+\infty$, then

$$V^{\mathrm{SC}}(x,y) - V^{\mathrm{SC}}(y,x) = (U(y)-U(x))^+ - (U(x)-U(y))^+ = U(y)-U(x).$$

Hence V^{SC} is induced by U. It remains to show that the family $(Q_\beta^{\mathrm{SC}})_\beta$ defined by (2.13) has rare transitions with rate function V^{SC}; that is, for all $(x,y)\in E^2$,

$$\lim_{\beta\to+\infty} -\beta^{-1}\ln Q_\beta^{\mathrm{SC}}(x,y) = \begin{cases} +\infty & \text{if } q^\star(x,y)=0, \\ 0 & \text{if } q^\star(x,y)>0 \text{ and } y=x, \\ (U(y)-U(x))^+ & \text{if } q^\star(x,y)>0 \text{ and } y\neq x. \end{cases}$$

Case 1. Assume that $q^\star(x,y)=0$. Then, by (v), we have $y\neq x$ and thus

$$-\beta^{-1}\ln Q_\beta^{\mathrm{SC}}(x,y) = -\beta^{-1}\ln q_\beta(x,y) + (U_\beta(y)-U_\beta(x))^+ \geqslant -\beta^{-1}\ln q_\beta(x,y).$$

Consequently, using (vi),

$$\lim_{\beta\to+\infty} -\beta^{-1}\ln Q_\beta^{\mathrm{SC}}(x,y) = +\infty.$$

Case 2. If $q^\star(x,y)>0$ and $y=x$, then

$$\exists\varepsilon\in(0,1),\ \exists\beta_0\geqslant 0,\ \forall\beta\geqslant\beta_0, \qquad q_\beta(x,x)\geqslant\varepsilon.$$

Therefore, for any $\beta\geqslant\beta_0$,

$$1 \geqslant Q_\beta^{\mathrm{SC}}(x,x) \geqslant 1 - \sum_{z\neq x} q_\beta(x,z) = q_\beta(x,x) \geqslant \varepsilon,$$

and thus $0 \leqslant -\beta^{-1}\ln Q_\beta^{\mathrm{SC}}(x,x) \leqslant -\beta^{-1}\ln\varepsilon$. It readily follows that

$$\lim_{\beta\to+\infty} -\beta^{-1}\ln Q_\beta^{\mathrm{SC}}(x,x) = 0.$$

Case 3. If $q^\star(x,y)>0$ and $y\neq x$, then

$$\lim_{\beta\to+\infty} -\beta^{-1}\ln Q_\beta^{\mathrm{SC}}(x,y) = \lim_{\beta\to+\infty} -\beta^{-1}\ln q_\beta(x,y) + (U(y)-U(x))^+,$$

and there exists $\varepsilon\in(0,1)$ such that $q_\beta(x,y)\geqslant\varepsilon$ for β large enough. Hence

$$\lim_{\beta\to+\infty} -\beta^{-1}\ln Q_\beta^{\mathrm{SC}}(x,y) = (U(y)-U(x))^+.$$

Acknowledgments. We thank the anonymous reviewers for their comments, which helped to improve the paper, and we thank Prof. Y. Boykov for his constructive suggestions as well as for pointing out to us the QPBO-based methods and references [5, 31, 34, 36, 49].

REFERENCES

[1] C. Bélisle, *Convergence theorems for a class of simulated annealing algorithms on* $\mathbb{R}^d$, J. Appl. Probab., 29 (1992), pp. 885–895.
[2] D. P. Bertsekas, *Network Optimization: Continuous and Discrete Models*, Athena Scientific, Nashua, NH, 1998.
[3] A. Blake, *Comparison of the efficiency of deterministic and stochastic algorithms for visual reconstruction*, IEEE Trans. Pattern Anal. Machine Intell., 11 (1989), pp. 2–12.
[4] A. Blake and A. Zisserman, *Visual Reconstruction*, The MIT Press, Cambridge, MA, 1987.
[5] E. Boros and P. Hammer, *Pseudo-Boolean optimization*, Discrete Appl. Math., 123 (2002), pp. 155–225.
[6] Y. Boykov, O. Veksler, and R. Zabih, *Fast approximate energy minimization via graph cuts*, IEEE Trans. Pattern Anal. Machine Intell., 23 (2001), pp. 1222–1239.
[7] O. Catoni, *Metropolis, simulated annealing, and iterated energy transformation algorithms: Theory and experiments*, J. Complexity, 12 (1996), pp. 595–623.
[8] O. Catoni, *Simulated annealing algorithms and Markov chains with rare transitions*, in Séminaire de probabilités XXXIII, Lecture Notes in Math. 1709, Springer, New York, 1999, pp. 69–119.
[9] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud, *Deterministic edge-preserving regularization in computed imaging*, IEEE Trans. Image Process., 6 (1997), pp. 298–311.
[10] S. Chen, R. Istepanian, and B. L. Luk, *Digital IIR filter design using adaptive simulated annealing*, Digital Signal Process., 11 (2001), pp. 241–251.
[11] T.-S. Chiang and Y. Chow, *On the convergence rate of annealing processes*, SIAM J. Control Optim., 26 (1988), pp. 1455–1470.
[12] C. Cot and O. Catoni, *Piecewise constant triangular cooling schedules for generalized simulated annealing algorithms*, Ann. Appl. Probab., 8 (1998), pp. 375–396.
[13] G. Davis, S. Mallat, and M. Avellaneda, *Adaptive greedy approximations*, Constr. Approx., 13 (1997), pp. 57–98.
[14] P. Del Moral and L. Miclo, *On the convergence and applications of generalized simulated annealing*, SIAM J. Control Optim., 37 (1999), pp. 1222–1250.
[15] A. H. Delaney and Y. Bresler, *Globally convergent edge-preserving regularized reconstruction: An application to limited-angle tomography*, IEEE Trans. Image Process., 7 (1998), pp. 204–221.
[16] G. Demoment, *Image reconstruction and restoration: Overview of common estimation structures and problems*, IEEE Trans. Acoust. Speech Signal Process., 37 (1989), pp. 2024–2036.
[17] O. François, *Global optimization with exploration/selection algorithms and simulated annealing*, Ann. Appl. Probab., 12 (2002), pp. 248–271.
[18] A. Frigerio and G. Grillo, *Simulated annealing with time-dependent energy function*, Math. Z., 213 (1993), pp. 97–116.
[19] D. Geiger and F. Girosi, *Parallel and deterministic algorithms from MRF's: Surface reconstruction*, IEEE Trans. Pattern Anal. Machine Intell., 13 (1991), pp. 401–412.
[20] S. B. Gelfand and S. K. Mitter, *Metropolis-type annealing algorithms for global optimization in* $\mathbb{R}^d$, SIAM J. Control Optim., 31 (1993), pp. 111–131.
[21] D. Geman, *Random fields and inverse problems in imaging*, in École d'été de Probabilités de Saint-Flour XVIII—1988, P. L. Hennequin, ed., Lecture Notes in Math. 1427, Springer, New York, 1990, pp. 117–193.
[22] D. Geman, S. Geman, C. Graffigne, and P. Dong, *Boundary detection by constrained optimization*, IEEE Trans. Pattern Anal. Machine Intell., 12 (1990), pp. 609–628.
[23] D. Geman and G. Reynolds, *Constrained restoration and the recovery of discontinuities*, IEEE Trans. Pattern Anal. Machine Intell., 14 (1992), pp. 367–383.
[24] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Trans. Pattern Anal. Machine Intell., 6 (1984), pp. 721–741.

[25] S. Geman and C. Graffigne, *Markov random field image models and their applications to computer vision*, in Proceedings of the International Congress of Mathematicians, Berkeley, CA, 1986, pp. 1496–1517.
[26] H. Haario and E. Saksman, *Simulated annealing process in general state space*, Adv. Appl. Probab., 23 (1991), pp. 866–893.
[27] B. Hajek, *Cooling schedules for optimal annealing*, Math. Oper. Res., 13 (1988), pp. 311–329.
[28] P. Hammer, P. Hansen, and B. Simeone, *Roof duality, complementation and persistency in quadratic 0-1 optimization*, Math. Program., 28 (1984), pp. 121–155.
[29] W. K. Hastings, *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika, 57 (1970), pp. 97–109.
[30] H. Ishikawa, *Exact optimization for Markov random fields with convex priors*, IEEE Trans. Pattern Anal. Machine Intell., 25 (2003), pp. 1333–1336.
[31] H. Ishikawa, *Higher-order clique reduction in binary graph cut*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, IEEE Press, Piscataway, NJ, pp. 2993–3000.
[32] Z. Kato and T.-C. Pong, *A Markov random field image segmentation model for color textured images*, Image Vision Comp., 24 (2006), pp. 1103–1114.
[33] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, *Optimization by simulated annealing*, Science, 220 (1983), pp. 671–680.
[34] P. Kohli, A. Shekhovtsov, C. Rother, V. Kolmogorov, and P. Torr, *On partial optimality in multi-label MRFs*, in Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 2008, ACM, New York, pp. 480–487.
[35] V. Kolmogorov, *Convergent tree-reweighted message passing for energy minimization*, IEEE Trans. Pattern Anal. Machine Intell., 28 (2006), pp. 1568–1583.
[36] V. Kolmogorov and C. Rother, *Minimizing non-submodular functions with graph cuts—A review*, IEEE Trans. Pattern Anal. Machine Intell., 29 (2007), pp. 1274–1279.
[37] V. Kolmogorov and R. Zabih, *What energy functions can be minimized via graph-cuts?*, IEEE Trans. Pattern Anal. Machine Intell., 26 (2004), pp. 147–159.
[38] Y. G. Leclerc, *Constructing stable descriptions for image partitioning*, Int. J. Comput. Vis., 3 (1989), pp. 73–102.
[39] S. Z. Li, *Markov Random Field Modeling in Computer Vision*, Springer, New York, 1995.
[40] S. Z. Li, *On discontinuity-adaptive smoothness priors in computer vision*, IEEE Trans. Pattern Anal. Machine Intell., 17 (1995), pp. 576–586.
[41] M. Locatelli, *Convergence properties of simulated annealing for continuous global optimization*, J. Appl. Probab., 33 (1996), pp. 1127–1140.
[42] M. Locatelli, *Simulated annealing algorithms for continuous global optimization: Convergence conditions*, J. Optim. Theory Appl., 104 (2000), pp. 121–133.
[43] M. Löwe, *Simulated annealing with time-dependent energy function via Sobolev inequalities*, Stochastic Process. Appl., 63 (1996), pp. 221–233.
[44] M. Nikolova, *Markovian reconstruction using a GNC approach*, IEEE Trans. Image Process., 8 (1999), pp. 1204–1220.
[45] M. Nikolova, *A variational approach to remove outliers and impulse noise*, J. Math. Imaging Vis., 20 (2004), pp. 99–120.
[46] M. Nikolova, *Analysis of the recovery of edges in images and signals by minimizing nonconvex regularized least-squares*, Multiscale Model. Simul., 4 (2005), pp. 960–991.
[47] M. Nikolova, J. Idier, and A. Mohammad-Djafari, *Inversion of large-support ill-posed linear operators using a piecewise Gaussian MRF*, IEEE Trans. Image Process., 7 (1998), pp. 571–585.
[48] M. Nikolova, M. K. Ng, S. Zhang, and W.-K. Ching, *Efficient reconstruction of piecewise constant images using nonsmooth nonconvex minimization*, SIAM J. Imaging Sci., 1 (2008), pp. 2–25.
[49] A. Raj, G. Singh, and R. Zabih, *MRF's for MRI's: Bayesian reconstruction of MR images via graph cuts*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New York, 2006, IEEE Press, Piscataway, NJ, pp. 1061–1068.
[50] A. Raj, G. Singh, R. Zabih, B. Kressler, Y. Wang, N. Schuff, and M. Weiner, *Bayesian parallel imaging with edge-preserving priors*, Magn. Reson. Med., 57 (2007), pp. 8–21.

[51] M. C. Robini, A. Lachal, and I. E. Magnin, *A stochastic continuation approach to piecewise constant reconstruction*, IEEE Trans. Image Process., 16 (2007), pp. 2576–2589.
[52] M. C. Robini and I. E. Magnin, *Stochastic nonlinear image restoration using the wavelet transform*, IEEE Trans. Image Process., 12 (2003), pp. 890–905.
[53] M. C. Robini, T. Rastello, and I. E. Magnin, *Simulated annealing, acceleration techniques and image restoration*, IEEE Trans. Image Process., 8 (1999), pp. 1374–1387.
[54] M. C. Robini, P.-J. Viverge, Y.-M. Zhu, and I. E. Magnin, *Edge-preserving image reconstruction with wavelet-domain edge continuation*, in Proceedings of the 6th International Conference on Image Analysis and Recognition, Halifax, Canada, Lecture Notes in Comput. Sci. 5627, Springer, New York, 2009, pp. 13–22.
[55] P. Rodríguez and B. Wohlberg, *Efficient minimization method for a generalized total variation functional*, IEEE Trans. Image Process., 18 (2009), pp. 322–332.
[56] C. Rother, V. Kolmogorov, V. Lempitsky, and M. Szummer, *Optimizing binary MRFs via extended roof duality*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, 2007, IEEE Press, Piscataway, NJ, pp. 1–8.
[57] C. Rother, S. Kumar, V. Kolmogorov, and A. Blake, *Digital tapestry*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, 2005, IEEE Press, Piscataway, NJ, pp. 589–596.
[58] R. H. Swendsen and J.-S. Wang, *Nonuniversal critical dynamics in Monte Carlo simulations*, Phys. Rev. Lett., 58 (1987), pp. 86–88.
[59] R. Szeliski and R. Zabih, *An experimental comparison of stereo algorithms*, in Vision Algorithms: Theory and Practice, Proceedings of the International Workshop on Vision Algorithms (Corfu, Greece, 1999), Lecture Notes in Comput. Sci. 1883, Springer, Berlin, 2000, pp. 1–19.
[60] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, *A comparative study of energy minimization methods for Markov random fields with smoothness-based priors*, IEEE Trans. Pattern Anal. Machine Intell., 30 (2008), pp. 1068–1080.
[61] A. Trouvé, *Massive Parallelization of Simulated Annealing*, Ph.D. thesis, 93 PA11 2030, Université Paris XI, Orsay, France, 1993 (in French).
[62] A. Trouvé, *Rough large deviation estimates for the optimal convergence speed exponent of generalized simulated annealing algorithms*, Ann. Inst. H. Poincaré Probab. Statist., 32 (1996), pp. 299–348.
[63] Y. Weiss and W. Freeman, *On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs*, IEEE Trans. Inform. Theory, 47 (2001), pp. 736–744.
[64] C. Yang, *Efficient stochastic algorithms on locally bounded image space*, Comput. Vision Graphics Image Process., 55 (1993), pp. 494–506.
[65] R. Yang, *Convergence of the simulated annealing algorithm for continuous global optimization*, J. Optim. Theory Appl., 104 (2000), pp. 691–716.
[66] J. Zhang and G. G. Hanauer, *The application of mean field theory to image motion estimation*, IEEE Trans. Image Process., 4 (1995), pp. 19–33.