Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

Stochastic annealing for nearest-neighbour point processes with application to object recognition

M N M van Lieshout

Department of Operations Reasearch, Statistics, and System Theory

BS-R9306 1993

Stochastic Annealing for Nearest-neighbour Point Processes with Application to Object Recognition

M. N. M. van Lieshout

Department of Mathematics and Computer Science, Free University
De Boelelaan 1081 a, 1081 HV Amsterdam, The Netherlands

CWI
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

Abstract: We study convergence in total variation of non-stationary Markov chains in continuous time and apply the results to image analysis problems such as object recognition and edge detection. Here the input is a grey-scale or binary image and the desired output is a graphical pattern in continuous space, such as a list of geometric objects or a line drawing. The natural prior models are Markov point processes found in stochastic geometry. We construct well defined spatial birth-and-death processes that converge weakly to the posterior distribution. A simulated annealing algorithm, involving a sequence of spatial birth-and-death processes is developed and shown to converge in total variation distance to a uniform distribution on the set of posterior mode solutions. The method is demonstrated on a tame example.

AMS Mathematics Subject Classification (1991 Revision): 60D05, 60J27, 62F15, 68T10.
Keywords & Phrases: convergence in total variation, Hough transform, likelihood ratio, maximum a posteriori estimation, nearest-neighbour Markov processes, spatial birth-and-death processes, stochastic annealing.

Introduction

In [6], Baddeley and Van Lieshout developed a statistical approach to the problem of object recognition in image analysis. Here a scene composed of possibly overlapping objects is observed subject to blur and noise, and the task is to determine the number of objects and locate them. Applications include document reading and robot vision. Their approach is formally similar to the well-known Bayesian formulation of image segmentation and classification problems due to Besag [7] and Geman & Geman [9], but differs in that discrete Markov random fields are replaced by continuous-space Markov spatial processes borrowed from stochastic geometry [26, 1].

The deterministic algorithms presented in [6] are strongly analogous to Besag's ICM method [7]. An obvious question is whether there is also an analogue of stochastic annealing [9, 10] in our context. The present paper explores the existence and convergence of certain spatial birth-and-death processes which are the counterparts of stochastic annealing algorithms (see also Winkler [28] for discrete time and [3]). Similar questions in autoradiography are apparently studied by Miller et al. [21].

The next section gives detailed background and notation. Section 2 introduces the class of nearest-neighbour Markov models that will be used as prior distribution. Convergence in total variation of inhomogeneous Markov processes is the subject of section 4. Section 5 studies existence and convergence of spatial birth-and-death processes on bounded subsets of $\mathbb{R}^d$, while section 6 treats the discrete case. The results of the previous sections are used in section 7 to develop analogues of simulated annealing in object recognition. Finally, section 8 gives a simple concrete example.

1 Preliminaries

This section gives an overview of notation and the main concepts that will be used in the sequel. For a more detailed description and examples we refer to [6].

1.1 Notation

The observed image $\mathbf{y}$ is digitized on a finite pixel lattice T ('image space') and for every $t \in T$, y_t denotes the value at pixel t. The values may be real, integer or binary or may belong to any arbitrary set. A typical choice is $\{0, 1, \ldots, 255\}$ for digital grey-level images.

The objects to be recognized are assumed to be representable by a finite number of real parameters that determine size, shape and location. Let U denote the set of possible vectors of parameter values ('object space'), so that a single point $u \in U$ represents an object $R(u) \subseteq T$. For example discs of fixed radius can be identified by their centre points, a line segment by position, orientation and length or an industrial robot can be specified by the position and orientation of the body and the attitude of each joint. We shall assume U is either a bounded Borel region ('continuous case') or a finite set of points in $\mathbb{R}^d$ ('discrete case'). Furthermore, U is equipped with its relative Borel σ-algebra and a reference measure $0 < \nu(U) < \infty$. We will take ν to be Lebesgue measure in the continuous case and counting measure in the discrete case.

An object configuration is an unordered list of objects

$$\mathbf{x} = \{x_1, \ldots, x_n\}, \ x_i \in U, \ i = 1, \ldots, n, \ n \geq 0.$$

The number of objects is variable and the empty list is allowed. Configuration $\mathbf{x}$ is often associated with its 'silhouette' scene $S(\mathbf{x}) = \bigcup R(x_i)$ in image space.

1.2 Independent noise models

The 'true' configuration $\mathbf{x}$ gives rise to the observed image $\mathbf{y}$ through a known probability distribution with density $f(\mathbf{y} \mid \mathbf{x})$. The stochastic models we use consist of a deterministic 'deformation' followed by random noise: every pattern $\mathbf{x}$ determines an ideal signal image $\theta^{(\mathbf{x})}$ in pixel space T; the values y_t are conditionally independent given $\mathbf{x}$ with the distribution of y_t depending only on the value of $\theta^{(\mathbf{x})}$ at that same pixel. Hence

$$f(\mathbf{y} \mid \mathbf{x}) = \prod_{t \in T} g(y_t \mid \theta^{(\mathbf{x})}(t))$$

where $g(\cdot|\theta)$ is a family of known probability densities (cf. Definition 1 in [6]). If the signal $\theta^{(\mathbf{x})}(t)$ at site t depends only on whether t belongs to the silhouette $S(\mathbf{x})$ or not, the model is called *blur-free*.

1.3 The Bayesian approach

In the Bayesian approach, the true image $\mathbf{x}$ is assumed to have been generated at random according to a prior probability distribution with density $p(\mathbf{x})$. The posterior probability for

$\mathbf{x}$ after observation of data image $\mathbf{y}$ is $p(\mathbf{x} \mid \mathbf{y}) \propto f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})$. The *Maximum A Posteriori* (MAP) estimator of $\mathbf{x}$ is

$$\tilde{\mathbf{x}} = \text{argmax}_{\mathbf{x}} f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x}). \tag{1}$$

The prior $p(\mathbf{x})$ can be viewed as a penalty assigned to the optimization; because of this interpretation MAP estimation is also known as *penalized maximum likelihood estimation*.

1.4 Connection with Hough transform

As (1) is an optimization problem over lists of variable length and additionally the prior is only known up to a normalizing constant, it is generally impossible to compute the MAP estimator analytically. Both the iterative methods in [6] and the alternatives based on spatial birth-and-death processes add and delete objects using log likelihood ratios that can be interpreted as the differences in 'goodness-of-fit' attained by altering the list.

Computing these log likelihood ratios usually involves only pixels 'local' to the altered object. More precisely, for blur-free independent noise models with $g(\cdot|\cdot) > 0$

$$L(\mathbf{x} \cup \{u\}\,;\ \mathbf{y}) - L(\mathbf{x}\,;\ \mathbf{y}) = \sum_{t \in R(u) \setminus S(\mathbf{x})} h(y_t, \theta_0, \theta_1) \tag{2}$$

where $L(\mathbf{x}\,;\ \mathbf{y}) = \log\ f(\mathbf{y} \mid \mathbf{x})$ and $h(y_t, \theta_0, \theta_1) = \log\ \frac{g(y_t \mid \theta_1)}{g(y_t \mid \theta_0)}$. The right hand side of (2) is a generalization of the *Hough transform* [17, 18] used in image processing for detecting simple objects [6, §4].

2 Prior object model

Here we introduce the class of Markov models used as prior distribution on object configurations. The models can be simulated by running a spatial birth-and-death process.

2.1 Nearest-neighbour Markov object processes

We need a prior model on the set of configurations $\Omega = \cup_{k=0}^{\infty} \mathcal{N}_k^{\#}(U)$,

$$\mathcal{N}_k^{\#}(U) = \{ \{x_1, \ldots, x_k\} \ : \ x_i \in U, \quad \text{all } x_i \text{ distinct } \},$$

defined by its density with respect to the Poisson process in the continuous case, or counting measure in the discrete case. Note that Ω is hereditary in the sense that $\mathbf{x} \in \Omega$ implies $\mathbf{y} \in \Omega$ for all subsets $\mathbf{y} \subseteq \mathbf{x}$. Moreover, in the continuous case, Ω has probability 1 under the Poisson model.

The appropriate prior models $p(\cdot)$ belong to the class of *nearest-neighbour Markov point processes* [1]. Their essential property is that objects in a given context interact only with their nearest neighbours. Formally, assume that for each $\mathbf{x} \in \Omega$ there is a symmetric relation $\sim_{\mathbf{x}}$ defined on $\mathbf{x}$. Related objects are called $\mathbf{x}$-*neighbours* and the $\mathbf{x}$- *neighbourhood* of any subset $\mathbf{y} \subseteq \mathbf{x}$ is defined by

$$N(\mathbf{y} \mid \mathbf{x}) = \{u \in \mathbf{x} \ : \ u \sim_{\mathbf{x}} v \text{ for some } v \in \mathbf{y}\}.$$

Finally, a subset $\mathbf{y} \subseteq \mathbf{x}$ is called a *clique* in $\mathbf{x}$ iff every pair of objects in $\mathbf{y}$ are $\mathbf{x}$-neighbours.

Definition 1 *A* **nearest-neighbour Markov object process** *with respect to* $\sim_{\mathbf{x}}$ *is a stochastic point process on* U *whose probability density* $p(\cdot)$ *satisfies*
(M1) $p(\mathbf{x}) > 0$ *implies* $p(\mathbf{y}) > 0$ *for all* $\mathbf{y} \subseteq \mathbf{x}$*;*
(M2) for each $\mathbf{x} \in \Omega$ *with* $p(\mathbf{x}) > 0$ *and each* $u \notin \mathbf{x}$*,*

$$\frac{p(\mathbf{x} \cup \{u\})}{p(\mathbf{x})}$$

depends only on u*,* $N(\{u\} \mid \mathbf{x} \cup \{u\})$ *and the restrictions of* $\sim_{\mathbf{x}}$*,* $\sim_{\mathbf{x} \cup \{u\}}$ *to* $N(\{u\} \mid \mathbf{x} \cup \{u\})$*.*

Remarks

1. Consider the case that $\sim_{\mathbf{x}}$ does not depend on $\mathbf{x}$. Then the definition is equivalent to that of a Markov random field in the discrete case and to a Ripley-Kelly process [26] in the continuous case.

2. A **Markov overlapping object process** [6] is a nearest-neighbour Markov object process with respect to the relation defined by

$$u \sim u' \Leftrightarrow R(u) \cap R(u') \neq \emptyset.$$

3. An example where $\sim_{\mathbf{x}}$ does depend on the context is for instance the **Dirichlet object process** for translation models where

$$u \sim u' \Leftrightarrow C(u \mid \mathbf{x}) \text{ and } C(u' \mid \mathbf{x}) \text{ share a common edge.}$$

   Here $C(u \mid \mathbf{x})$ denotes the Voronoi cell of u in configuration $\mathbf{x}$.

An equivalent definition in terms of iteractions between objects can be obtained from an analogue of the Hammersley-Clifford theorem.

Theorem 1 *(***Baddeley-Møller, 1989***)*
Under certain regularity assumptions [1, C(1)-C(2)] the point process with density $p(\cdot)$ is a nearest-neighbour Markov object process with respect to $\sim_{\mathbf{x}}$ if and only if there exist interaction functions [1, Definition 4.11] Φ such that

$$p(\mathbf{x}) = \prod_{\mathbf{y} \subseteq \mathbf{x}} \Phi(\mathbf{y} \mid \mathbf{x})$$

for all $\mathbf{x} \in \Omega$.

Usually we will consider only pairwise interactions

$$p(\mathbf{x}) = \alpha \prod_{i=1}^{n} \beta(x_i) \prod_{i<j; x_i \sim_{\mathbf{x}} x_j} q(x_i, x_j)$$

To enforce (M2) we require the relation $\sim_{\mathbf{x}}$ to satisfy

$$\chi(\mathbf{y} \mid \mathbf{z}) \neq \chi(\mathbf{y} \mid \mathbf{z} \cup \{u\}) \text{ implies } \mathbf{y} \subseteq N(\{u\} \mid \mathbf{z} \cup \{u\});$$

for all $\mathbf{y} \subseteq \mathbf{x} \in \Omega$, $\mathbf{z} \not\ni u \in U$, where χ denotes the clique indicator function. Furthermore $\alpha > 0$ is a normalizing constant, $\beta : U \to (0, \infty)$ a strictly positive measurable function and $q : U \times U \to [0, \infty)$ a measurable function making $p(\cdot)$ integrate up to 1 and satisfying one of the following conditions.

(a) $q(\cdot, \cdot) > 0$;

(b) if $q(x_i, x_j) > 0$ for all $x_i \sim_{\mathbf{x}} x_j \in \mathbf{x}$ then $q(x_i, x_j) > 0$ for all $x_i, x_j \in \mathbf{x}$;

(c) $x_i \sim_{\mathbf{x}} x_j$ implies $x_i \sim_{\mathbf{x} \cup \{u\}} x_j$ for all $u \notin \mathbf{x}$.

Any of (a)-(c) implies (M1), hence $p(\cdot)$ is a Markov function. The special case that β and q are constant is called a *Strauss process.*

To enforce mutual inhibition between objects we require $q(x_i, x_j) < 1$ for all $x_i, x_j \in \mathbf{x}$ with $x_i \sim_{\mathbf{x}} x_j$.

2.2 Spatial birth-and-death processes

A Markov object process may arise as the equilibrium distribution of a *spatial birth-and-death process* [1, 24, 26]. In our context this is a continuous time, pure jump Markov process with states in Ω, with the property that the only transitions are the birth of a new object (a transition from $\mathbf{x}$ to $\mathbf{x} \cup \{u\}$) or the death of an existing one (transition from $\mathbf{x}$ to $\mathbf{x} \setminus \{u\}$). The process is said to have birth rate $B(\mathbf{x}, F)$ and death rate $D(\mathbf{x}, u)$, for measurable $F \subseteq U$ and $u \in U$, when it can be described as follows:

- given the state $\mathbf{x}$ at time t, the waiting time to the first transition after time t is independent of the history of the process and exponentially distributed with parameter $D(\mathbf{x}) + B(\mathbf{x})$, where
$$\begin{array}{rcl} D(\mathbf{x}) & = & \sum_{x_i \in \mathbf{x}} D(\mathbf{x} \setminus \{x_i\}, x_i) \\ B(\mathbf{x}) & = & B(\mathbf{x}, U). \end{array}$$
- the next transition is a birth with probability $B(\mathbf{x})/[D(\mathbf{x}) + B(\mathbf{x})]$ and a death with probability $D(\mathbf{x})/[D(\mathbf{x}) + B(\mathbf{x})]$.
- given that the next transition is a birth, the new point belongs to the measurable subset F with probability $B(\mathbf{x}, F)/B(\mathbf{x})$.
- given that the next transition is a death, x_i is deleted from $\mathbf{x}$ with probability $D(\mathbf{x} \setminus \{x_i\}, x_i)/D(\mathbf{x})$.

If $B(\mathbf{x}, \cdot)$ is absolutely continuous with respect to the reference measure ν on U, its Radon-Nikodym density will be denoted by $b(\mathbf{x}, \cdot)$.

However a given set of rates does not necessarily define a unique spatial-birth-and-death process; conditions on $B(\cdot)$ and $D(\cdot)$ have to be imposed to avoid explosion, i.e. an infinite number of state transitions can occur in a finite time.

3 Convergence of inhomogeneous Markov chains

3.1 Definitions

Let μ and ν be probability measures on a common measurable space $(\mathcal{S}, \mathcal{A})$. Their **total variation distance** is defined as the maximal difference in mass on measurable subsets $A \in \mathcal{A}$

$$|| \mu - \nu || = \sup_{A \in \mathcal{A}} | \mu(A) - \nu(A) |.$$

If $|\mathcal{S}| < \infty$

$$|| \mu - \nu || = \frac{1}{2} \sum_{i \in \mathcal{S}} | \mu(i) - \nu(i) |.$$

Similarly in the continuous case, if both μ and ν are absolutely continuous with respect to some measure m with Radon-Nikodym derivatives f_μ and f_ν,

$$|| \mu - \nu || = \frac{1}{2} \int_{\mathcal{S}} | f_\mu(s) - f_\nu(s) | \, m(ds).$$

Definition 2 *For a transition probability function (stochastic matrix) $P(\cdot, \cdot)$ on $(\mathcal{S}, \mathcal{A})$,* **Dobrushin's contraction coefficient** *$c(P)$ is defined by*

$$c(P) = \sup_{x,y \in \mathcal{S}} || P(x, \cdot) - P(y, \cdot) ||$$

We list some properties that will be used in the sequel (see Dobrushin [8, section 3]).

Lemma 2 *Let Λ be the set of all probability measures on $(\mathcal{S}, \mathcal{A})$. Then for all transition probability functions P and Q and for all $\mu, \nu \in \Lambda$ the following hold*

(i) $c(P) \leq 1$;

(ii) $c(PQ) \leq c(P)\, c(Q)$;

(iii) $c(P) = \sup_{\mu \neq \nu \in \Lambda} \frac{|| \mu P - \lambda P ||}{|| \mu - \lambda ||}$;

(iv) $|| \mu P - \nu P || \leq c(P) \, || \mu - \nu ||$;

(v) *If $\mathcal{S}$ is finite, $c(P) = 1 - \rho(P)$, where $\rho(P)$ denotes the* **ergodic coefficient** *of P*

$$\rho(P) = \min_{x,y \in I} \sum_{k \in I} P(x, k) \wedge P(y, k),$$

($\wedge$ denoting minimum).

3.2 Limit theorems

The main theorems of this section state sufficient conditions under which a sequence of Markov chains converges in total variation to a well-defined limit.

Theorem 3 *Let $(X_t)_{t\geq 0}$ be a non-stationary Markov chain in continuous time on a finite state space I, with transition rates defined by a sequence $(R_n)_{n\in\mathbb{N}}$, such that $R_{n(t)}$ is in force at time t, with $n(t) \nearrow \infty$ as $t \to \infty$. Assume that for each $n \in \mathbb{N}$ the 'rate in = rate out' equations have a unique solution μ_n and*

$$\begin{array}{ll} (C) & \displaystyle\sum_{n=1}^{\infty} ||\mu_n - \mu_{n+1}|| < \infty \\ (D) & c(P_{tt'}) \to 0 \ \text{ as } t' \to \infty \ \text{ for all } t \geq 0 \end{array}$$

where $P_{tt'}(i,j) = \mathbb{P}(X_{t'} = j \mid X_t = i)$. Then $\mu_\infty = \lim \mu_n$ exists and $\nu P_{0t} \to \mu_\infty$ in total variation as $t \to \infty$, uniformly in the initial distribution ν.

Theorem 4 *Let $(X_t)_{t\geq 0}$ be a non-stationary Markov process on a measurable space $(\mathcal{S}, \mathcal{A})$, defined by a sequence of transition semi-groups $(Q^n)_{n\in\mathbb{N}}$, such that $Q^{n(t)}$ is in force at time t, with $n(t) \nearrow \infty$ as $t \to \infty$. Assume that for each $n \in \mathbb{N}$, Q^n has an invariant measure μ_n, i.e.*

$$\int_{\mathcal{S}} Q_t^n(x, F)\, \mu_n(dx) = \mu_n(F)$$

for all $F \in \mathcal{A}$ and $t \geq 0$. Assume moreover that the following hold

$$\begin{array}{ll} (C) & \displaystyle\sum_{n=1}^{\infty} ||\mu_n - \mu_{n+1}|| < \infty \\ (D) & c(P_{tt'}) \to 0 \ \text{ as } t' \to \infty \ \text{ for all } t \geq 0 \end{array}$$

where $P_{tt'}(x, F) = \mathbb{P}(X_{t'} \in F \mid X_t = x)$. Then $\mu_\infty = \lim \mu_n$ exists and $\nu P_{0t} \to \mu_\infty$ in total variation as $t \to \infty$, uniformly in the initial distribution ν.

Proof: (both cases)
Condition (C) implies that (μ_n) is a Cauchy sequence in $||\cdot||$ and hence converges in total variation to μ_∞, say.
Define $t_k = \inf\{t \,:\, n(t) \geq k\}$ and choose $0 \leq t < t_{n(t)+1} < t' < \infty$. Then

$$\begin{aligned} \mu_\infty P_{tt'} - \mu_\infty &= (\mu_\infty - \mu_{n(t)}) P_{tt'} + \mu_{n(t)} P_{tt'} - \mu_\infty \\ &= (\mu_\infty - \mu_{n(t)}) P_{tt'} + \mu_{n(t)} P_{tt_{n(t)+1}} P_{t_{n(t)+1}t'} - \mu_\infty. \end{aligned}$$

Since $\mu_{n(t)}$ is an invariant measure one sees that

$$\begin{aligned} \mu_\infty P_{tt'} - \mu_\infty &= (\mu_\infty - \mu_{n(t)}) P_{tt'} + \mu_{n(t)} P_{t_{n(t)+1}t'} - \mu_\infty \\ &= (\mu_\infty - \mu_{n(t)}) P_{tt'} + \sum_{k=n(t)}^{n(t')-1} (\mu_k - \mu_{k+1}) P_{t_{k+1}t'} + \mu_{n(t')} - \mu_\infty. \end{aligned}$$

Hence

$$\begin{aligned}
\| \mu_\infty P_{tt'} - \mu_\infty \| &\le \| \mu_\infty - \mu_{n(t)} \| \, c(P_{tt'}) + \sum_{k=n(t)}^{n(t')-1} \| \mu_k - \mu_{k+1} \| \, c(P_{t_{k+1}t'}) + \| \mu_{n(t')} - \mu_\infty \| \\
&\le 2 \sup_{k \ge n(t)} \| \mu_k - \mu_\infty \| + \sum_{k=n(t)}^{\infty} \| \mu_k - \mu_{k+1} \| \\
&\quad \to 0 \ (t \to \infty).
\end{aligned}$$

Let $\epsilon > 0$. Choose t such that $\| \mu_\infty P_{tt'} - \mu_\infty \| < \epsilon/2$ for all $t' > t_{n(t)+1}$. Next observe that

$$\begin{aligned}
\| \nu P_{0t'} - \mu_\infty \| &= \| (\nu P_{0t} - \mu_\infty) P_{tt'} + \mu_\infty P_{tt'} - \mu_\infty \| \\
&\le \| \nu P_{0t} - \mu_\infty \| \, c(P_{tt'}) + \| \mu_\infty P_{tt'} - \mu_\infty \| \\
&\le c(P_{tt'}) + \| \mu_\infty P_{tt'} - \mu_\infty \|.
\end{aligned}$$

Use condition (D) to choose t' such that $c(P_{tt'}) < \epsilon/4$. Summarizing, we obtain

$$\| \nu P_{0t} - \mu_\infty \| \to 0 \ \text{ uniformly in } \nu \ (t \to \infty).$$

□

A sufficient condition for (D) is given by the next result. It is easier to work with, since only stationary Markov chains have to be considered.

Lemma 5 *Use the same notation as in the previous Theorems. If $c(P_{t_n t_{n+1}}) \le 1 - 1/n$ for all $n \ge 2$, the Dobrushin condition (D) holds.*

Proof: Write $P_n = P_{t_n t_{n+1}}$. Then

$$-\log c(P_n) \ge 1 - c(P_n) \ge \frac{1}{n}.$$

Thus

$$-\sum_{n=2}^{\infty} \log c(P_n) \ge \sum_{n=2}^{\infty} \frac{1}{n} = \infty$$

or equivalently

$$\prod_{n=2}^{N} c(P_n) \to 0 \ (N \to \infty).$$

For $t < t'$. Then

$$\begin{aligned}
c(P_{tt'}) &= c(P_{tt_{n(t)+1}} P_{t_{n(t)+1} t_{n(t)+2}} \cdots P_{t_{n(t')} t'}) \\
&\le c(P_{tt_{n(t)+1}}) \left[\prod_{i=n(t)+1}^{n(t')-1} c(P_i) \right] c(P_{t_{n(t')} t'}) \\
&\le \prod_{i=n(t)+1}^{n(t')-1} c(P_i) \to 0 \ (t' \to \infty).
\end{aligned}$$

□

4 Birth-and-death processes on bounded regions

Here we study existence and convergence of spatial birth-and-death processes on bounded subsets of $\mathbb{R}^d$. The results are used to construct birth-and-death processes that converge in distribution to the (modified) posterior distribution.

4.1 Motivation

To recover the underlying object configuration from blurred and noisy observations, equations (1) must be solved. To sharpen peaks in posterior probability, set

$$p_H(\mathbf{x} \mid \mathbf{y}) \propto \{f(\mathbf{y} \mid \mathbf{x})\ p(\mathbf{x})\}^{1/H}.$$

For small $H > 0$, configurations with large posterior density are favoured, while others are suppressed. Indeed, if object space U is discrete $p_H(\cdot \mid \mathbf{y})$ converges pointwise to a uniform distribution on the set of MAP solutions as H tends to zero.

For each fixed H we can construct a spatial birth-and-death process with equilibrium distribution $p_H(\cdot \mid \mathbf{y})$. Our proposal therefore is to use a stochastic annealing algorithm that simulates these processes consecutively with H gradually dropping to zero.

In the superficially similar context of image segmentation, a simulated annealing algorithm was developed by Geman and Geman [9]. However, the Markov processes involved are rather different. Since in segmentation problems both object and image space are finite pixel grids, a discrete time Markov chain changing each pixel label in turn suffices.

4.2 Theorem

Preston [24] gave sufficient conditions under which there exists a unique spatial birth-and-death process with given rates solving Kolmogorov's backward equations

$$\frac{d}{dt}\mathbb{P}(X_t \in F | X_0 = \mathbf{x}) = -[B(x) + D(x)]\ \mathbb{P}(X_t \in F | X_0 = \mathbf{x}) + \int \mathbb{P}(X_t \in F | X_0 = \mathbf{z}) R(\mathbf{x}, d\mathbf{z})$$

with $R(\mathbf{x}, F)$ being the total rate from pattern $\mathbf{x}$ into F. For a given process (X_t) he also found conditions for the existence of a unique invariant probability measure and convergence in distribution (i.e. convergence of $\mathbb{P}(X_t \in F | X_0 = \mathbf{x})$). Baddeley and Møller [1] combined these two theorems to obtain the following result.

Theorem 6 *For each $n = 0, 1, \ldots$ define*

$$\kappa_n = \sup_{n(\mathbf{x})=n} B(\mathbf{x})$$

and

$$\delta_n = \inf_{n(\mathbf{x})=n} D(\mathbf{x}).$$

Assume $\delta_n > 0$ for all $n \geq 1$. If either (a) *that $\kappa_n = 0$ for all sufficiently large $n \geq 0$,* or (b) *that $\kappa_n > 0$ for all $n \geq 0$ and both the following hold:*

$$\sum_{n=1}^{\infty} \frac{\kappa_0 \cdots \kappa_{n-1}}{\delta_1 \cdots \delta_n} < \infty$$
$$\sum_{n=1}^{\infty} \frac{\delta_1 \cdots \delta_n}{\kappa_1 \cdots \kappa_n} = \infty$$

then there exists a unique spatial birth-and-death process for which $B(\cdot)$ and $D(\cdot)$ are the transition rates; this process has a unique equilibrium distribution to which it converges in distribution from any initial state.

A slightly stronger result given by Møller [22] includes the case $\kappa_0 = 0$, $\kappa_n > 0$ for all $(n \geq 1)$, $\sum_{n=2}^{\infty} \frac{\kappa_1 \cdots \kappa_{n-1}}{\delta_1 \cdots \delta_n} < \infty$ and $\sum_{n=1}^{\infty} \frac{\delta_1 \cdots \delta_n}{\kappa_1 \cdots \kappa_n} = \infty$, still assuming all δ_n positive for $n \geq 1$.

4.3 Construction

In this section we shall construct well-defined spatial birth-and-death processes that converge in distribution to $p_H(\cdot|\mathbf{y})$. Consider any blur-free independent noise model (§1.3) with $g(\cdot|\cdot) > 0$ and a nearest-neighbour Markov object prior. The former assumption is needed so that the class $K = \{\mathbf{x} \ : \ f(\mathbf{y} \mid \mathbf{x})\ p(\mathbf{x}) > 0\}$ is *hereditary* [26], that is if a configuration $\mathbf{x}$ belongs to K the same is true for all its subsets $\mathbf{y} \subseteq \mathbf{x}$. For some fixed $k \in [0,1]$ set

$$b_H(\mathbf{x}, u) = \begin{cases} \left(\frac{f(\mathbf{y} \mid \mathbf{x} \cup \{u\})\, p(\mathbf{x} \cup \{u\})}{f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})}\right)^{k/H} & \text{if } f(\mathbf{y} \mid \mathbf{x})\ p(\mathbf{x}) > 0 \\ 0 & \text{if } f(\mathbf{y} \mid \mathbf{x})\ p(\mathbf{x}) = 0 \end{cases} \tag{3}$$

for $u \notin \mathbf{x}$ and death rate

$$D_H(\mathbf{x} \setminus \{u\}, u) = \begin{cases} \left(\frac{f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})}{f(\mathbf{y} \mid \mathbf{x} \setminus \{u\})\, p(\mathbf{x} \setminus \{u\})}\right)^{\frac{k-1}{H}} & \text{if } f(\mathbf{y} \mid \mathbf{x})\ p(\mathbf{x}) > 0 \\ \delta'_n / n & \text{if } f(\mathbf{y} \mid \mathbf{x})\ p(\mathbf{x}) = 0,\ n(\mathbf{x}) = n \end{cases} \tag{4}$$

Here $\delta'_n = \inf \{\sum_{x_i \in \mathbf{x}} D_H(\mathbf{x} \setminus \{x_i\}, x_i) \mid f(\mathbf{y} \mid \mathbf{x})\ p(\mathbf{x}) > 0,\ n(\mathbf{x}) = n\}$. By convention, the infimum of the empty set equals ∞. Note that by this definition $\delta'_n = \delta_n$, where δ_n is defined as in Theorem 6. The boundary cases $k = 0$ ('constant birth rate') and $k = 1$ ('constant death rate') are well-known in spatial statistics, to obtain realizations of a point process. It is argued by several authors [25] that the constant death rate procedure should be preferred, as the other method tends to add an unlikely object, and then deletes it immediately.

Note that these transition rates do not depend on the normalizing constant in p. Moreover the 'detailed balance' equations

$$b_H(\mathbf{x}, u)\ p_H(\mathbf{x} \mid \mathbf{y}) = D_H(\mathbf{x}, u)\ p_H(\mathbf{x} \cup \{u\} \mid \mathbf{y}) \tag{5}$$

are satisfied whenever $f(\mathbf{y} \mid \mathbf{x} \cup \{u\})\ p(\mathbf{x} \cup \{u\}) > 0$. Given a spatial birth-and-death process with rates satisfying (5), Ripley [25] remarked that $p_H(\cdot|\mathbf{y})$ is necessarily the density of a unique invariant probability measure for the process. In general however, there is no guarantee of existence, nor of convergence to the proper limit distribution.

Corollary 7 *Let $\mathbf{y}$ and $H > 0$ be fixed. For any blur-free independent noise model with $g(\cdot|\cdot) > 0$, and any nearest-neighbour Markov object process $p(\cdot)$ with uniformly bounded likelihood ratios*

$$\frac{p(\mathbf{x} \cup \{u\})}{p(\mathbf{x})} \leq \beta < \infty$$

there exists a unique spatial birth-and-death process for which (3) and (4) are the transition rates. The process has unique equilibrium distribution $p_H(\cdot \mid \mathbf{y})$ and it converges in distribution to $p_H(\cdot \mid \mathbf{y})$ from any initial state.

Proof: We will prove the following properties:

1. $\delta_n > 0$, for $n \geq 1$;
2. if $\kappa_{n_0} = 0$ for some $n_0 \geq 1$, then $\kappa_n = 0 \ \forall n \geq n_0$;
3. if $\kappa_n > 0$ for all n, then condition (b) of Theorem 6 holds.

Property 1: Use the representation of the log likelihood ratio as a generalized Hough transform (2). Since T is finite, we have upper and lower bounds on the goodness of fit, say $\mid h(y_t, \theta_0, \theta_1) \mid \leq a$ for all t. Hence

$$\mid L(\mathbf{x} \cup \{u\} \ ; \ \mathbf{y}) - L(\mathbf{x} \ ; \ \mathbf{y}) \mid \leq a \ n(R(u)) \leq a \ n(T)$$

where n denotes 'area' or number of pixels. For $p(\cdot)$ we have by assumption $\frac{p(\mathbf{x} \cup \{u\})}{p(\mathbf{x})} \leq \beta$ If $p(\mathbf{x}) > 0$ this implies that

$$\begin{aligned} D_H(\mathbf{x} \setminus \{u\}, u) &= \left(\frac{f(\mathbf{y} \mid \mathbf{x})}{f(\mathbf{y} \mid \mathbf{x} \setminus \{u\})} \frac{p(\mathbf{x})}{p(\mathbf{x} \setminus \{u\})} \right)^{\frac{k-1}{H}} \\ &\geq \exp\left[\frac{k-1}{H} \left(\mid L(\mathbf{x} \setminus \{u\} \ ; \ \mathbf{y}) - L(\mathbf{x} \ ; \ \mathbf{y}) \mid + \log \beta \right) \right] \\ &\geq \exp\left[\frac{k-1}{H} \left(a \ n(T) + \log \beta \right) \right] \\ &=: \delta > 0. \end{aligned}$$

Suppose $p(\mathbf{x}) = 0$. If $p(\mathbf{z}) = 0$ for all $\mathbf{z}$ with $n(\mathbf{z}) = n(\mathbf{x})$, then $D_H(\mathbf{x} \setminus \{u\}, u) = \infty \geq \delta$. Otherwise $n(\mathbf{x}) \ D_H(\mathbf{x} \setminus \{u\}, u) = \inf \{D_H(\mathbf{z}) \mid n(\mathbf{z}) = n(\mathbf{x}), \ p(\mathbf{z}) > 0\}$. By the argument above $D_H(\mathbf{z} \setminus \{z_i\}, z_i) \geq \delta$ for all such $\mathbf{z}$ and $z_i \in \mathbf{z}$. Hence $D_H(\mathbf{z}) \geq \delta \ n(\mathbf{x})$ and $D_H(\mathbf{x} \setminus \{u\}, u) \geq \delta$. Therefore $D_H(\mathbf{x}) \geq \delta \ n(\mathbf{x})$ for all patterns $\mathbf{x}$, and hence $\delta_n \geq \delta \ n > 0$ for $n \geq 1$.

Property 2: Use the fact that $K = \{\mathbf{x} \ : \ f(\mathbf{y} \mid \mathbf{x}) \ p(\mathbf{x}) > 0\}$ is hereditary.

Property 3: The birth rates are also bounded. For $p(\mathbf{x}) > 0$

$$\begin{aligned} b_H(\mathbf{x}, u) &= \left(\frac{f(\mathbf{y} \mid \mathbf{x} \cup \{u\})}{f(\mathbf{y} \mid \mathbf{x})} \frac{p(\mathbf{x} \cup \{u\})}{p(\mathbf{x})} \right)^{k/H} \\ &\leq \exp\left[\frac{k}{H} \left(a \ n(T) + \log \beta \right) \right] \\ &=: \kappa > 0. \end{aligned}$$

For $p(\mathbf{x})=0$, again $b_H(\mathbf{x},u)=0\leq\kappa$. Hence $B_H(\mathbf{x})\leq\kappa\,\nu(U)$ and so $\kappa_n\leq\kappa\,\nu(U)$.

Using these bounds, one obtains

$$\frac{\kappa_0\cdots\kappa_{n-1}}{\delta_1\cdots\delta_n}\leq\frac{\kappa^n\,\nu(U)^n}{n!\,\delta^n}$$

Since $\nu(U)$ is finite by assumption, the first assertion follows. Similarly

$$\frac{\delta_1\cdots\delta_n}{\kappa_1\cdots\kappa_n}\geq\frac{n!\,\delta^n}{\kappa^n\,\nu(U)^n}$$

which does not converge to zero as $n\to\infty$. The Corollary is proved if we combine these properties with Theorem 6. $\square$

5 Birth-and-death processes on finite spaces

In this section we assume that object space U is a discrete subset of $\mathbb{R}^d$, say $U = \{u_1, \ldots, u_M\}$ equipped with counting measure.

5.1 Construction

Recall that any Markov chain on a finite state space is uniquely defined by its transition rates. We will need the following result about the long run behaviour [23, Chapter 7].

Lemma 8 *Let $(X)_{t \geq 0}$ be an irreducible Markov chain on a finite state space I, specified by its transition rates $R(i,j)$, $i \neq j$. If π is a probability measure on I satisfying the 'detailed balance' equations*

$$R(i,j)\,\pi(i) = R(j,i)\,\pi(j) \tag{6}$$

then (X_t) is time reversible and has unique limiting distribution π.

Hence, if a spatial birth-and-death process defined by birth rates $b(\cdot,\cdot)$ and death rates $D(\cdot,\cdot)$ is irreducible and π is a probability measure on $\mathcal{P}(U)$ satisfying (6), then the process is time reversible and has unique limiting distribution π.

5.2 Simulation

Returning to the object recognition context (section 1), we focus attention on the constant death rate procedure with death rates $D(\cdot,\cdot) = 1$ and birth rates defined by (3) with $k = 1$. Assume $f(\mathbf{y} \mid \mathbf{x}) > 0$ for all x. Then the class $K = \{\mathbf{x} \ : \ f(\mathbf{y} \mid \mathbf{x})\ p(\mathbf{x}) > 0\}$ is hereditary, due to the Markov property of $p(\cdot)$. It follows that the process restricted to K is irreducible. The total birth rate is simply

$$B_H(\mathbf{x}) = \sum_j b_H(\mathbf{x}, u_j)$$

and the total death rate is $D_H(\mathbf{x}) = n(\mathbf{x})$, the number of elements in list $\mathbf{x}$.

To simulate the birth-death process for fixed data $\mathbf{y}$ and $H > 0$, note that the process can be represented as a sequence $(X^{(k)}, T^{(k)})$, $k = 1, 2, \ldots$ of random variables. Here $X^{(k)}$ are the successive states and $T^{(k)}$ is the sojourn time in state $X^{(k)}$. Given $X^{(k)} = \mathbf{x}^{(k)}$, time $T^{(k)}$ is exponentially distributed with mean $1/(n(\mathbf{x}^{(k)}) + B_H(\mathbf{x}^{(k)}))$, independent of other sojourn times and of past states. The next state transition is a death with probability $n(\mathbf{x}^{(k)})/(n(\mathbf{x}^{(k)}) + B_H(\mathbf{x}^{(k)}))$, obtained by deleting one of the existing points with equal probability; otherwise the transition is a birth generated by choosing one of the points $u_j \notin \mathbf{x}^{(k)}$ with probability

$$\frac{b_H(\mathbf{x}^{(k)}, u_j)}{B_H(\mathbf{x}^{(k)})}$$

and adding u_j to the state. This representation is implemented directly by the following algorithm.

Algorithm 1 *Select an initial state* $\mathbf{x}^{(0)}$. *To simulate from the equilibrium distribution, the process described above is run for a 'large' time period* C, *and we take* $\mathbf{x}^{(K)}$ *where*

$$K = \min\{k = 0, 1, 2, \ldots \mid \sum_{i=0}^{k} t^{(i)} > C\}.$$

5.3 Practical considerations

Since the birth rate $b_H(\mathbf{x}, u)$ is an exponential function of the Hough transform (2), it tends to have sharp peaks as a function of u when H is small or when $\mathbf{x}$ is far from a MAP solution. There is then a high probability that the next transition will add a new object u at one of the locations that is close to maximal for $b_H(\mathbf{x}, u)$. This implies that naive rejection sampling methods are not a workable alternative to Algorithm 1; since these methods generate a putative new object uniformly, many would be rejected and the waiting times would be unacceptably long.

To avoid numerical instability or overflow one could first compute

$$c_H(\mathbf{x}, u) = \log\ b_H(\mathbf{x}, u) = H^{-1}\left(L(\mathbf{x} \cup \{u\}\ ;\ \mathbf{y}) - L(\mathbf{x}\ ;\ \mathbf{y}) + \log\ \frac{p(\mathbf{x} \cup \{u\})}{p(\mathbf{x})}\right)$$

for each $u = u_j$, to find

$$m(\mathbf{x}) = \max_j\ c_H(\mathbf{x}, u_j)$$

then to compute

$$B_H(\mathbf{x}) = e^{m(\mathbf{x})}\ \sum_j \exp\left[c_H(\mathbf{x}, u_j) - m(\mathbf{x})\right]$$

where the exponential term lies between 0 and 1. The conditional probability of a birth at u_j is computable from the same summands,

$$\frac{b_H(\mathbf{x}, u_j)}{B_H(\mathbf{x})} = \exp\left[c_H(\mathbf{x}, u_j) - m(\mathbf{x})\right] \exp\left[m(\mathbf{x}) - \log B_H(\mathbf{x})\right]$$

For small H and $\mathbf{x}^{(k)}$ far from optimal, the following approximate algorithm would be more efficient.

Algorithm 2 *Let* $\epsilon > 0$ *be given.*

- *Compute* $n = n(\mathbf{x}^{(k)})$ *and for each* u_j *compute* $c_j = \log b_H(\mathbf{x}^{(k)}, u_j)$;
- *Find* $m(\mathbf{x}^{(k)}) = \max_j c_j$;
- *Take* $v = v(H, \mathbf{x}^{(k)}) \geq 0$ *so large that*

$$\nu(U) \exp\left[m(\mathbf{x}^{(k)}) - v\right] \leq \epsilon;$$

- *Rank all the values* c_j *for* $m(\mathbf{x}^{(k)}) - v < c_j \leq m(\mathbf{x}^{(k)})$ *to give a list of values* $m(\mathbf{x}^{(k)}) = m_1 > m_2 > \ldots > m_L > m(\mathbf{x}^{(k)}) - v$, *and compute the 'multiplicities'*

$$k_\ell = \nu(\{u_j : c_j = m_\ell\});$$

- *Approximate* $B_H(\mathbf{x}^{(k)})$ *by*

$$B = \sum_{\ell=1}^{L} k_\ell e^{m_\ell};$$

- *Proceed as in Algorithm 1, except that to generate a birth, the procedure is to choose one of the values* $\ell = 1, \ldots, L$ *with probability* $k_\ell e^{m_\ell}/B$*; then one of the* k_ℓ *points* u_j *such that* $c_j = m_\ell$ *is chosen with equal probability yielding* $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} \cup \{u_j\}$.

The choice of v is arranged so that B is an approximation to $B_H(\mathbf{x})$ with error at most ϵ:

$$0 \leq B_H(\mathbf{x}) - B = \sum_{c_j \leq m(\mathbf{x}) - v} e^{c_j} \leq \nu(U) \exp\left[m(\mathbf{x}) - v\right] \leq \epsilon.$$

For very small H, when $\mathbf{x}^{(k)}$ is not close to the MAP solution, Algorithm 2 amounts to choosing **at random** one of the u_j that maximizes the conditional posterior likelihood ratio $c_H(\mathbf{x}, u)$. This is in a sense comparable to the behaviour of the steepest ascent algorithm presented in [6], but we note that this description ceases to hold as the solution approaches optimality.

6 Stochastic annealing for point processes

Here we present an alternative method for solving the MAP equations (1), using the results of section 3.

6.1 The summability condition

Lemma 9 *Let $H_n \searrow 0$ $(n \to \infty)$ and consider the sequence of H_n-modified posterior distributions with densities*

$$p_{H_n}(\mathbf{x} \mid \mathbf{y}) \propto \{f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})\}^{1/H_n}$$

with respect to reference measure m on Ω (counting measure or law of Poisson process). Assume that $m(\mathcal{M}) > 0$, where $\mathcal{M}$ denotes the set of solutions to the MAP equations (1). Then the sequence p_{H_n} converges in total variation to a uniform distribution on $\mathcal{M}$. Moreover the sequence satisfies condition (C).

Proof:

Since $m(\mathcal{M}) > 0$, $\exists\, \mathbf{x}^{\#}$ attaining the maximum. Denote

$$Z_n = \int_\Omega \left(\frac{f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})}{f(\mathbf{y} \mid \mathbf{x}^{\#})\, p(\mathbf{x}^{\#})} \right)^{1/H_n} m(d\mathbf{x}).$$

We have

$$\begin{aligned}
\lim_{n\to\infty} Z_n &= \lim_{n\to\infty} \int_\Omega \left(\frac{f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})}{f(\mathbf{y} \mid \mathbf{x}^{\#})\, p(\mathbf{x}^{\#})} \right)^{1/H_n} m(d\mathbf{x}) \\
&= \int_\Omega \lim_{n\to\infty} \left(\frac{f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})}{f(\mathbf{y} \mid \mathbf{x}^{\#})\, p(\mathbf{x}^{\#})} \right)^{1/H_n} m(d\mathbf{x}) \\
&= \int_\Omega 1\{\mathbf{x} \in \mathcal{M}\} m(d\mathbf{x}) \\
&= m(\mathcal{M}).
\end{aligned}$$

The second equation follows by the dominated convergence theorem. Moreover, $Z_n \downarrow m(\mathcal{M})$, $\sup Z_n = Z_1 \leq m(\Omega) < \infty$. It is easily seen that

$$l_n(\mathbf{x}) := \left(\frac{f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})}{f(\mathbf{y} \mid \mathbf{x}^{\#})\, p(\mathbf{x}^{\#})} \right)^{1/H_n} \to 1\{\mathbf{x} \in \mathcal{M}\}, \quad n \to \infty,$$

proving the first assertion.

To prove condition (C), suppress the dependence on $\mathbf{x}$ and consider

$$\begin{aligned}
\left| \frac{l_n}{Z_n} - \frac{l_{n+1}}{Z_{n+1}} \right| &= (Z_n Z_{n+1})^{-1} \mid l_n Z_{n+1} - l_{n+1} Z_n \mid \\
&\leq (Z_n Z_{n+1})^{-1} \; \{ \, l_{n+1} | \, Z_{n+1} - Z_n \mid \, + \, Z_{n+1} \mid l_{n+1} - l_n \mid \} \\
&\leq \frac{1}{m(\mathcal{M})^2} \{ Z_{n+1}(l_n - l_{n+1}) + l_{n+1}(Z_n - Z_{n+1}) \} \\
&\leq \frac{m(\Omega)}{m(\mathcal{M})^2} \{ l_n - l_{n+1} + Z_n - Z_{n+1} \}
\end{aligned}$$

Therefore

$$\begin{aligned}
\sum_{n=1}^{N-1} \int_\Omega \mid \frac{l_n(\mathbf{x})}{Z_n} - \frac{l_{n+1}(\mathbf{x})}{Z_{n+1}} \mid m(d\mathbf{x}) &\leq \sum_{n=1}^{N-1} \frac{m(\Omega)}{m(\mathcal{M})^2} \int_\Omega \{l_n(\mathbf{x}) - l_{n+1}(\mathbf{x}) + Z_n - Z_{n+1}\}\, m(d\mathbf{x}) \\
&= \frac{m(\Omega)}{m(\mathcal{M})^2} \int_\Omega \sum_{n=1}^{N-1} \{l_n(\mathbf{x}) - l_{n+1}(\mathbf{x}) + Z_n - Z_{n+1}\}\, m(d\mathbf{x}) \\
&= \frac{m(\Omega)}{m(\mathcal{M})^2} \int_\Omega \{l_1(\mathbf{x}) - l_N(\mathbf{x}) + Z_1 - Z_N\}\, m(d\mathbf{x}) \\
&= \frac{m(\Omega)}{m(\mathcal{M})^2} \,(1 + m(\Omega))\,(Z_1 - Z_N)
\end{aligned}$$

Letting $N \to \infty$ the right hand side converges to $\frac{m(\Omega)}{m(\mathcal{M})^2}\,(1 + m(\Omega))\,(Z_1 - m(\mathcal{M})) < \infty$. □

Note that the assumption $m(\mathcal{M})$ is automatically satisfied in the discrete case. The assumption is needed; if $m(\mathcal{M}) = 0$ the sequence of modified posterior distributions will not converge in total variation (cf. [16]).

6.2 The Dobrushin condition

From now on let f be a blur-free independent noise model with $g(\cdot|\cdot) > 0$. Again, let $H_n \searrow 0$ and consider the family $(X^{(n)})_{n\in\mathbb{N}}$ of spatial birth-and-death processes on $K = \{\mathbf{x} : f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x}) > 0\}$ defined by (3) and (4). As K is closed and irreducible, the processes are well-defined and converge to the unique limit $p_H(\cdot \mid \mathbf{y})$ (see sections 4 and 5).

Recall the following result by Møller [22], a generalization of earlier work by Lotwick and Silverman [20].

Theorem 10 *Let X_t be a spatial birth-and-death process and define κ_n, δ_n as in Theorem 6. Assume moreover that $\delta_n > 0$ for all $n \geq 1$ and $\kappa_n = 0$ for all $n > n_0$ (condition (a)). Then for all fixed $t_0 > 0$*

$$\sup_{\mathbf{x},\mathbf{y}} \| \, P_t(\mathbf{x}, \cdot) - P_t(\mathbf{y}, \cdot) \, \| \leq 2\,(1 - K(t_0))^{\frac{t}{t_0} - 1}$$

for all $t > t_0$.

Here

$$\tilde{\delta}_m = \min_{i,j:i+j=m} \delta_i + \delta_j; \;\; \tilde{\kappa}_m = \max_{i,j:i+j=m} \kappa_i + \kappa_j; \;\; \tilde{\alpha}_m = \tilde{\delta}_m + \tilde{\kappa}_m;$$

$$K_1(j,t) = \left[1 - e^{-\tilde{\alpha}_j t}\right] \frac{\tilde{\delta}_j}{\tilde{\alpha}_j}; \;\; K_0(n, t_0) = \prod_{j=1}^{n} K_1(j, \frac{t_0}{n}); \;\; K(t_0) = \min_{n \leq n_0} K_0(n, t_0);$$

Therefore, by Lemma 5 we can construct an annealing schedule satisfying condition (D) by requiring

$$2(1 - K(t_0))^{\frac{t}{t_0} - 1} \leq 1 - \frac{1}{n}$$

or equivalently

$$t \geq t_0 \left(1 + \frac{\log(\frac{1}{2}\,(1-\frac{1}{n}))}{\log(1-K(t_0))}\right).$$

By Lemma 9 condition (C) also holds. Finally, by Theorem 3 the sequence of birth-and-death processes constructed this way converges in total variation to a uniform distribution on the set of global maxima of the posterior distribution, regardless of the initial state.

6.3 Remarks and extensions

In the finite case $|U| < \infty$, generalizations to diffusing objects are possible. Set

$$M(\mathbf{x}, x_i, u)$$

for the configuration obtained from $\mathbf{x}$ by replacing x_i by u. The set of $u \in U$ for which this operation is allowed is denoted by $Q(\mathbf{x}, x_i)$. Typically it consists of unoccupied objects close but not identical to x_i.

Suppose the diffusion rates are also powers of the log likelihood ratios

$$c_H(\mathbf{x}, x_i, u) = \left\{ \frac{f(\mathbf{y} \mid M(\mathbf{x}, x_i, u))\; p(M(\mathbf{x}, x_i, u))}{f(\mathbf{y} \mid \mathbf{x})\; p(\mathbf{x})} \right\}^{k/H}. \tag{7}$$

If detailed balance is required to hold as well, necessarily $k = \frac{1}{2}$.

The following additional notation is needed.

$$\gamma_i = \max_{|\mathbf{x}|=i} C(\mathbf{x}).$$

$$\tilde{\delta}_m = \min_{i,j:i+j=m} \delta_i + \delta_j;\ \tilde{\kappa}_m = \max_{i,j:i+j=m} \kappa_i + \kappa_j;\ \tilde{\gamma}_m = \max_{i,j:i+j=m} \gamma_i + \gamma_j;\ \tilde{\alpha}_m = \tilde{\delta}_m + \tilde{\kappa}_m + \tilde{\gamma};$$

$$K_1'(j,t) = \left[1 - e^{-\tilde{\alpha}_j t}\right] \frac{\tilde{\delta}_j}{\tilde{\alpha}_j};\ K_0'(n, t_0) = \prod_{j=1}^{n} K_1'(j, \frac{t_0}{n});\ K'(t_0) = \min_{n \leq n_0} K_0'(n, t_0);$$

Theorem 11 *The Markov chain on K with rates given by (3), (4) and (7) is irreducible and satisfies the detailed balance equations. It is time reversible and has unique equilibrium distribution $p_H(\cdot|\mathbf{y})$. Moreover for all fixed $t_0 > 0$*

$$\sup_{\mathbf{x},\mathbf{y}} ||\ P_t(\mathbf{x}, \cdot) - P_t(\mathbf{y}, \cdot)\ || \leq 2\ (1 - K'(t_0))^{\frac{t}{t_0} - 1}$$

for all $t > t_0$.

Proof: Every configuration $\mathbf{x}$ can be reached from every other configuration $\mathbf{z}$ by first deleting all points in $\mathbf{x} \setminus \mathbf{z}$ and then adding $\mathbf{z} \setminus \mathbf{x}$, since all transition rates are strictly positive. Thus the Markov chain is irreducible. Since the detailed balance equations hold by construction, the first part of the result follows from lemma 8. The second part can be derived by coupling arguments (see Appendix). □

7 Discussion

We conclude this paper with a simple practical example and some comparison between deterministic and stochastic algorithms.

7.1 Example

In [6] we studied in detail a simple example where a scene composed of discs with fixed radius was observed after addition of white Gaussian noise (Figure 1). A Strauss process served as prior distribution on disc configurations.

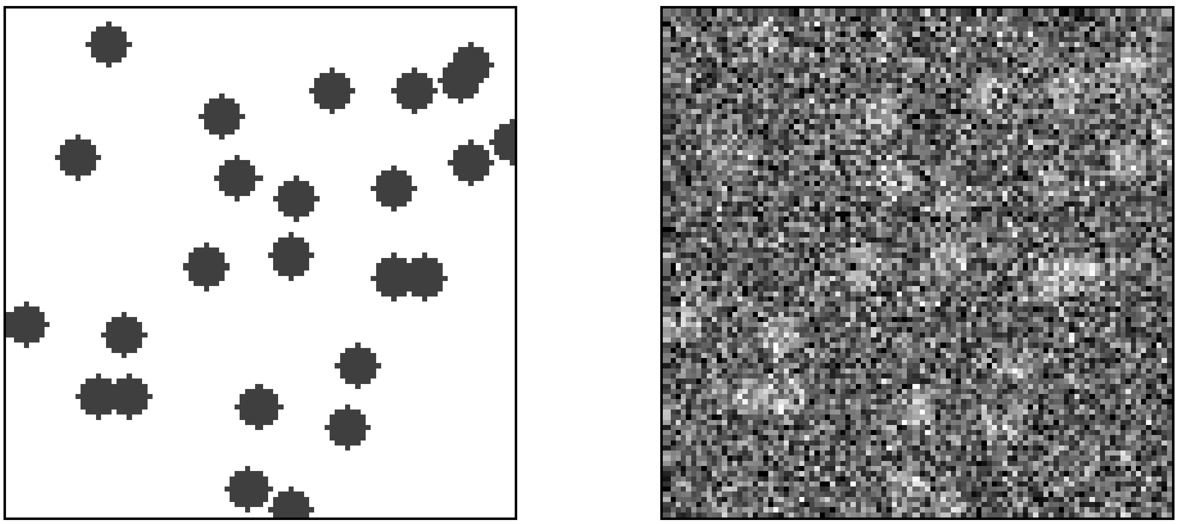

Figure 1: Binary silhouette scene of discs with radius 4 and realization from Gaussian model with $\sigma = 50$, $\theta_1 = 150$ and $\theta_0 = 100$ digitized on a 98×98 square grid.

To obtain an approximate MAP solution, in principle a sequence of spatial birth-and-death processes should be run. However, since the posterior distribution is highly peaked it is sufficient to sample at a fixed 'low' temperature (see also [11]). A reconstruction obtained by running the constant death rate procedure at $H = 1$ is given in Figure 2. Using constant birth rate instead was found to behave worse. The latter method tends to add an unlikely object and immediately deletes it again. This confirms experience reported in the literature ([22, 25]).

Typical runs of the constant death rate method are illustrated in Figure 3, where the Δ_2 distance [2] to the 'true' pattern is graphed against time. Starting from an empty scene, objects are immediately added to form a plausible reconstruction followed by deletion and immediate readding of one of the objects. Note that the results obtained this way are comparable to steepest ascent reconstructions [6].

In contrast to ICM, simulated annealing results in a global maximum, regardless of the initial state. Experiments with several initial patterns are in accordance with the theory, in that similar reconstructions were obtained.

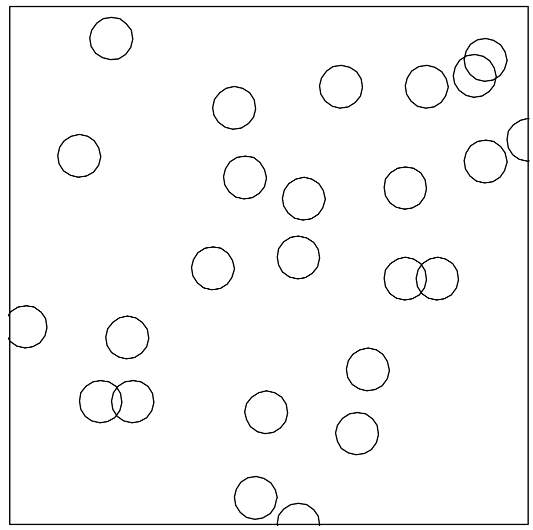

Figure 2: MAP reconstruction using a Strauss prior with $\beta = .0025$ and $\gamma = .25$.

7.2 Concluding remarks

Here the model was specified completely. In reality there will be unknown physical parameters that have to be estimated. This can be done either from training data, or if none is available by alternating reconstruction and parameter estimation steps ([6, section 7]).

The computational effort per transition is mainly in sampling from $b_H(\mathbf{x}, \cdot)/B(\mathbf{x})$. For finite parameter spaces, this can be done naively by one scan through object space, but more efficient strategies should be used. The performance of the method on realistic image data is closely related and will be the topic of further research.

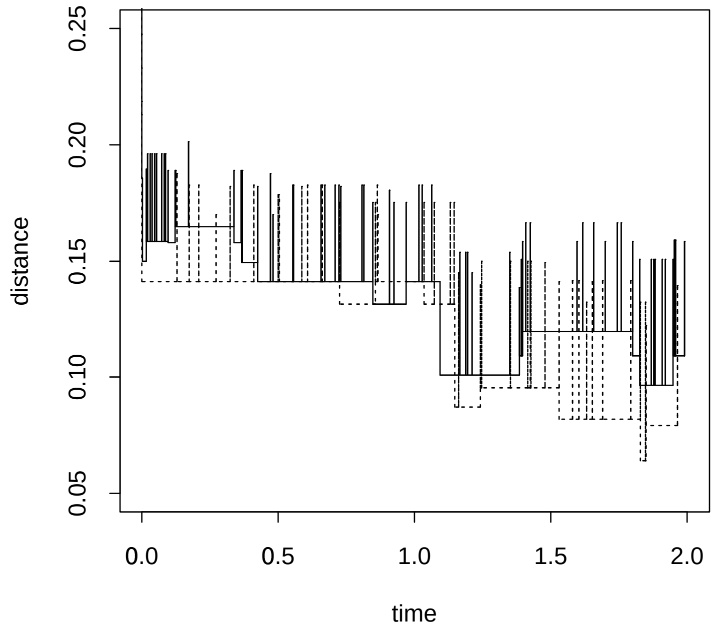

Figure 3: Delta-2 distance between reconstructed and true pattern as a function of time. The cutoff value is 4.

Acknowledgements

I am very grateful to my research supervisor Prof A J Baddeley for his support and suggestions, to A G Steenbeek for technical assistance and to Prof R D Gill and Dr J Møller for helpful feedback.

References

[1] Baddeley, A. and Møller, J. (1989) Nearest-neighbour Markov point processes and random sets. *International Statistical Review* **57**, 89–121.

[2] Baddeley, A.J. (1992) An error metric for binary images. In: Förstner, W. and Ruwiedel, S. (eds.), *Robust Computer Vision*, 59 – 78, Wichmann, Karlsruhe.

[3] Baddeley, A.J. and Van Lieshout, M.N.M (1991) Recognition of overlapping objects using Markov spatial models. CWI Report BS-R9109, March 1991.

[4] Baddeley, A.J. and Van Lieshout, M.N.M. (1992) ICM for object recognition. In: Dodge, Y. and Whittaker, J. (eds.), *Computational Statistics*, Volume 2, 271 – 286, Physica/Springer, Heidelberg-New York.

[5] Baddeley, A.J. and Van Lieshout, M.N.M. (1992) Object recognition using Markov spatial processes. In: *Proceedings 11th IAPR International Conference on Pattern Recognition*, Volume B, 136 – 139, IEEE Computer Society Press, Los Alamitos.

[6] Baddeley, A.J. and Van Lieshout, M.N.M. (1992) Recognition of overlapping objects using Markov spatial processes. Part I: Deterministic algorithms. *Manuscript submitted for publication.*

[7] Besag, J. (1986) On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society, Series B* **48**, 259 – 302.

[8] Dobrushin, R.L (1956) Central limit theorem for non-stationary Markov chains I, II. *Theory of Probability and its Applications* **1**, 65 – 80, 329 – 383.

[9] Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721 – 741.

[10] Geman, D. (1990) Random fields and inverse problems in imaging. In: *Ecole d'ete de probabilites de Saint-Flour XVIII - 1988; Lecture Notes in Mathematics* **1427**. Berlin: Springer-Verlag.

[11] Geman, D., Geman, S., Graffigne, C. and Dong, P. (1990) Boundary detection by constrained optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**, 609 – 628.

[12] Grenander, U. (1976) *Lectures on Pattern Theory.* Vol. 1: *Pattern Synthesis.* New York-Berlin: Springer-Verlag, Applied Mathematical Sciences vol. 18.

[13] Grenander, U. (1978) *Lectures on Pattern Theory.* Vol. 2: *Pattern Analysis.* New York-Berlin: Springer-Verlag, Applied Mathematical Sciences vol. 24.

[14] Grenander, U. (1981) *Lectures on Pattern Theory.* Vol. 3: *Regular Structures.* New York-Berlin: Springer-Verlag, Applied Mathematical Sciences vol. 33.

[15] Grenander, U. and Keenan, D.M. (1989) A computer experiment in pattern theory. *Communications in Statistics - Stochastic Models* **5**, 531 – 553.

[16] Haario, H. and Saksman, E. (1991) Simulated annealing process in general state space. *Advances in Applied Probability* **23**, 866 – 893.

Biometrika **57**, 97 – 109.

[17] Hough, P.V.C. (1962) Method and means for recognizing complex patterns. US Patent 3,069,654.

[18] Illingworth, J. and Kittler, J. (1988) A survey of the Hough transform. *Computer Vision, Graphics and Image Processing* **44**, 87 – 116.

[19] Ten Kate, T.K., Van Balen, R., Smeulders, A.W.M., Groen, F.C.A. and Den Boer, G.A. (1990) SCILAIM: a multi-level interactive image processing environment. *Pattern Recognition Letters* **11**, 429 – 441.

[20] Lotwick, H.W. and Silverman, B.W. (1981) Convergence of spatial birth-and-death processes. *Mathematical Proceedings of the Cambridge Philosophical Society* **90**, 155 – 165.

[21] Miller, M. et al. (1991) Automated segmentation of biological shapes in electron microscopic autoradiography. In: Davidson, F and Goutsias, J. (Eds), 25th Annual Conference on Information Sciences and Systems.

[22] Møller, J. (1989) On the rate of convergence of spatial birth-and-death processes. *Annals of the Institute of Statistical Mathematics* **41**, 565 – 581.

[23] Parzen, E. (1962) *Stochastic processes.* San Francisco: Holden-Day.

[24] Preston, C.J. (1977) Spatial birth-and-death processes. *Bulletin of the International Statistical Institute* **46**, 371 – 391.

[25] Ripley, B.D. (1977) Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 172 – 212.

[26] Ripley, B.D. and Kelly, F.P. (1977) Markov point processes. *Journal of the London Mathematical Society* **15**, 188 – 192.

[27] Stoyan, D., Kendall, W.S. and Mecke J. (1987) *Stochastic geometry and its applications.* Chichester: John Wiley and Sons.

[28] Winkler, G. (1990) An ergodic L^2 theorem for simulated annealing in Bayesian image reconstruction. *Journal of Applied Probability* **28**, 779 – 791.

Appendix

Markov chains on finite spaces

We will need the following results from Markov chain theory (Chung, 1967).

Lemma 12 *Let Q be a finite square matrix satisfying*

$$\begin{array}{lll} (i) & q_{ij} \geq 0 & \forall i \neq j \\ (ii) & q_{ii} = -\sum_{j \neq i} q_{ij} & \forall i \end{array}$$

Then there exist unique transition matrices

$$P_t = e^{Qt} = \sum_{n=0}^{\infty} \frac{Q^n t^n}{n!}, \; t \geq 0$$

so that the associated Markov chain has rates Q. They form the unique solution to the Kolmogorov differential equations ($t \geq 0$)

$$\begin{cases} z'_{ij}(t) & = & \sum_k z_{ik}(t) q_{kj} \\ z'_{ij}(t) & = & \sum_k q_{ik} z_{kj}(t) \end{cases}$$

In other words, any Markov chain on a finite state space is uniquely specified by its transition rates $R(i,j) = q_{ij}$, $i \neq j$.

A Markov chain is called irreducible, if all states can be reached from one another. This means that for all states $i \neq j$ there exists a sequence $i = i_1, i_2, \cdots, i_n = j$ such that

$$R(i_1, i_2) \cdots R(i_{n-1}, i_n) > 0.$$

For irreducible Markov chains the long run fraction of time spent in each state can be found be solving a system of linear equations.

Lemma 13 *Let (X_t) be a Markov chain on a finite state space with transition matrices P_t, $t \geq 0$. If X is irreducible*

$$\lim_{t \to \infty} P_t(i,j) = \pi(j)$$

exists for all i, j and is independent of the initial state. The limit vector π is the unique solution of the 'rate in = rate out' equations

$$\begin{cases} \pi Q & = & 0 \\ \sum_i \pi(i) & = & 1 \end{cases}$$

where Q is the transition rates matrix $q_{ij} = \frac{d}{dt} p'_t(i,j)|_{t=0}$.

Lemma 14 *Let X be an irreducible Markov chain with finite state space I, specified by its transition rates $R(i,j)$. If π is a probability measure on I satisfying the 'detailed balance' equations*

$$R(i,j)\pi(i) = R(j,i)\pi(j)$$

then X is time reversible and has unique limiting distribution π.

Rate of convergence

The following is a summary of results by Lotwick and Silverman (1981) and Møller (1989). For notation see the main text.

Let X_t and Y_t be two independent simple birth-and-death processes and set $Z_t = (X_t, Y_t)$. Suppose $Z_0 = (m, n)$ with $m + n > 0$. Then

$$\mathbb{P}(\text{ the first transition in } Z \text{ occurs before time } t \text{ and is a death} \mid Z_0 = (m,n)) =$$

$$\left[1 - e^{-(\alpha_m + \alpha_n)t}\right] \frac{\delta_m + \delta_n}{\alpha_m + \alpha_n} \geq K_1(m + n, t).$$

To see the last inequality, set

$$f(x, y) = [1 - e^{-x-y}] \frac{y}{x + y}$$

where $x = (\kappa_m + \kappa_n)t \geq 0$ and $y = (\delta_m + \delta_n)t > 0$. The partial derivatives of f are

$$\frac{\partial f}{\partial x}(x, y) = \frac{y}{(x + y)^2} [(1 + x + y)e^{-x-y} - 1];$$

$$\frac{\partial f}{\partial y}(x, y) = \frac{y}{x + y} e^{-x-y} + \frac{x}{(x + y)^2} (1 - e^{-x-y}) > 0.$$

Set $g(z) = (1 + z)e^{-z} - 1$ on $(0, \infty)$. Since $g'(z) = -ze^{-1} < 0$, $\frac{\partial f}{\partial x}(x, y) < 0$. Therefore a lower bound can be obtained by taking x as large and y as small as possible.

Also:

$$\mathbb{P}(Z_t \text{ hits } (0,0) \text{ before time } t_0 \mid Z_0 = (m,n)) \geq$$

$$\mathbb{P}(\text{ first } m + n \text{ transitions occur before } t_0 \text{ and all are deaths} \mid (m,n)) \geq$$

$$\mathbb{P}(\text{ first } m + n \text{ sojourn times are all } \leq \tfrac{t_0}{m+n} \text{ and all these transitions are deaths} \mid (m,n)) \geq$$

$$K_0(m + n, t_0)$$

If the number of objects is bounded above by n_0 (e.g if the state space is finite), then $K(t_0) = \min_{n \leq n_0} K_0(n, t_0) > 0$ and

$$\mathbb{P}(Z_t \text{ hits } (0,0) \text{ before time } t_0) \geq K(t_0)$$

for all initial distributions.

Lemma: For all $k \in \mathbb{N}$

$$1 - \mathbb{P}(Z_t \text{ hits } (0,0) \text{ before } kt_0) \leq (1 - K(t_0))^k.$$

The proof is by induction. For $k = 1$ the assertion is trivially correct. Next assume the lemma holds for some $k \in \mathbb{N}$. Then, abreviating 'hits' by $\Uparrow$,

$$1 - \mathbb{P}(Z_t \Uparrow (0,0) \text{ before } (k+1)t_0) =$$

$$\int \mathbb{P}(Z_t \not\to (0,0) \text{ in } (t_0,(k+1)t_0) \,|z) * \mathbb{P}(Z_t \not\to (0,0) \text{ before } t_0 \text{ and } Z_{t_0} \in dz) \leq$$

$$(1-K(t_0))^k \int \mathbb{P}(Z_t \not\to (0,0) \text{ before } t_0 \text{ and } Z_{t_0} \in dz) =$$

$$(1-K(t_0))^k \, \mathbb{P}(Z_t \not\to (0,0) \text{ before } t_0) \leq (1-K(t_0))^{k+1}$$

which proves the Lemma.

Now, for arbitrary $t > t_0$,

$$\mathbb{P}(Z_t \not\to (0,0) \text{ before } t) \leq$$

$$\mathbb{P}(Z_t \not\to (0,0) \text{ before } \lfloor \frac{t}{t_0} \rfloor t_0) \leq$$

$$(1-K(t_0))^{\lfloor \frac{t}{t_0} \rfloor} \leq (1-K(t_0))^{\frac{t}{t_0}-1}$$

Next, consider independent **general** spatial birth-and-death processes X_t and Y_t and assume that the number of objects is bounded. Moreover, let the initial distributions be λ (concentrated on configurations with at most n_0 points) and the equilibrium measure π respectively. Since state $(0,0)$ is discrete, coupling applies. Define

$$U_t = \begin{cases} X_t, & t < \tau \\ Y_t, & t \geq \tau \end{cases}$$

where $\tau = \inf\{t > 0 \ : \ X_t = Y_t = 0\}$.

Then τ is a stopping time and $U_t \stackrel{d}{=} X_t$. Hence for every measurable A,

$$\begin{aligned} \mathbb{P}_\lambda(X_t \in A) - \pi(A) &= \mathbb{P}_\lambda(U_t \in A) - \pi(A) \\ &= \mathbb{P}_\lambda(X_t \in A \ ; \ t < \tau) + \mathbb{P}_\lambda(Y_t \in A \ ; \ t \geq \tau) - \pi(A) \\ &\leq \mathbb{P}_\lambda(X_t \in A \ ; \ t < \tau) + \mathbb{P}_\lambda(Y_t \in A) - \pi(A) \\ &= \mathbb{P}_\lambda(X_t \in A \ ; \ t < \tau) \leq \mathbb{P}(t < \tau). \end{aligned}$$

and similarly

$$\begin{aligned} \pi(A) - \mathbb{P}_\lambda(X_t \in A) &= \pi(A) - \mathbb{P}_\lambda(U_t \in A) \\ &= \pi(A) - \mathbb{P}_\lambda(X_t \in A \ ; \ t < \tau) - \mathbb{P}_\lambda(Y_t \in A \ ; \ t \geq \tau) \\ &= \mathbb{P}_\lambda(Y_t \in A \ ; \ t < \tau) - \mathbb{P}_\lambda(X_t \in A \ ; \ t < \tau) \\ &\leq \mathbb{P}_\lambda(Y_t \in A \ ; \ t < \tau) \leq \mathbb{P}_\lambda(t < \tau). \end{aligned}$$

Thus

$$|| \, \pi - \mathbb{P}_\lambda(X_t \in \cdot) \, ||_{TV} \leq \mathbb{P}_\lambda(t < \tau)$$

and the theorem is proved.