Statistics and Probability Letters 113 (2016) 1–6

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Geometric ergodicity of Rao and Teh's algorithm for homogeneous Markov jump processes

Błażej Miasojedow*, Wojciech Niemiro

Institute of Applied Mathematics and Mechanics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland

ARTICLE INFO

Article history:
Received 2 December 2015
Received in revised form 8 February 2016
Accepted 8 February 2016
Available online 20 February 2016

Keywords:
Continuous time Markov processes
MCMC
Hidden Markov models
Geometric ergodicity
Drift condition
Small set

ABSTRACT

Rao and Teh (2013) introduced an efficient MCMC algorithm for sampling from the posterior distribution of a hidden Markov jump process. The algorithm is based on the idea of sampling virtual jumps. In the present paper we show that the Markov chain generated by Rao and Teh's algorithm is geometrically ergodic. To this end we establish a geometric drift condition towards a small set. We work under the assumption that the parameters of the hidden process are known and the goal is to restore its trajectory.

1. Introduction

Markov jump processes (MJP) are natural extension of Markov chains to continuous time. They are widely applied in modelling of the phenomena of chemical, biological, economic and other sciences.

In many applications it is necessary to consider a situation where the trajectory of a MJP is not observed directly, only partial and noisy observations are available. Typically, the posterior distribution over trajectories is then analytically intractable. In the literature there exist several approaches to the above mentioned problem: based on sampling (Boys et al., 2008; El-Hay et al., 2008; Fan and Shelton, 2008; Golightly and Wilkinson, 2011, 2014; Golightly et al., 2015; Nodelman et al., 2002; Rao and Teh, 2013, 2012), and also based on numerical approximations. To the best of our knowledge the most general efficient method for a finite state space is that proposed by Rao and Teh (2013), and extended to a more general class of continuous time discrete systems in Rao and Teh (2012). Although the method proposed by Fearnhead and Sherlock (2006), after a minor modification, can be used to sample exactly from the posterior distribution of a homogeneous MJP, their algorithm involves calculating matrix exponentials, which is computationally expensive.

In the present paper we establish geometric ergodicity of Rao and Teh's algorithm for homogeneous MJPs. Geometric ergodicity is a key property of Markov chains which implies Central Limit Theorem for sample averages.

Note that in practice the parameters of the hidden MJP may be unknown and have to be estimated. Then the Rao and Teh's algorithm can be combined with an additional step (Gibbs or Metropolis–Hastings) updating these parameters, according to some posterior distribution. Such extended versions of the Rao and Teh's algorithm are not considered in our paper. We assume that the probability law of a hidden MJP is known.

* Corresponding author.
E-mail addresses: bmia@mimuw.edu.pl (B. Miasojedow), wniem@mimuw.edu.pl (W. Niemiro).

http://dx.doi.org/10.1016/j.spl.2016.02.002
0167-7152/

The rest of the paper is organized as follows. In Section 2 we briefly introduce hidden Markov jump processes, next in Section 3 we recall the Rao and Teh's algorithm. The main result is proved in Section 4.

2. Hidden Markov jump processes

Consider a continuous-time homogeneous Markov process $\{X(t),\, t^{\min} \le t \le t^{\max}\}$ on a finite state space $\mathcal{S}$. Its probability law is defined via the initial distribution $\nu(s) = \mathbb{P}(X(t^{\min}) = s)$ and the transition intensities

$$Q(s, s') = \lim_{h \to 0} \frac{1}{h} \mathbb{P}(X(t+h) = s' | X(t) = s)$$

for $s, s' \in \mathcal{S},\ s \neq s'$. Let $Q(s) = \sum_{s' \neq s} Q(s, s')$ denote the intensity of leaving state s. For definiteness, assume that X has right-continuous trajectories. We say X is a Markov jump process (MJP).

Suppose that process X cannot be directly observed but we can observe some random quantity Y with probability distribution $L(Y|X)$. Let us say Y is the evidence and L is the likelihood. The problem is to restore the hidden trajectory of X given Y. From the Bayesian perspective, the goal is to compute/approximate/sample from the posterior

$$p(X|Y) \propto p(X) L(Y|X).$$

Function L, transition probabilities Q and initial distribution ν are assumed to be known. To get the explicit form of posterior distribution we consider a typical form of noisy observation. Assume that the trajectory $X([t^{\min}, t^{\max}])$ is observed independently at k deterministic time points with some random errors. Formally, we observe $Y = (Y_1, \ldots, Y_k)$ where

$$L(Y|X) = \prod_{j=1}^{k} L_j(Y_j | X(t_j^{\text{obs}})), \tag{1}$$

for some fixed known points $t^{\min} \le t_1^{\text{obs}} < \cdots < t_k^{\text{obs}} \le t^{\max}$.

3. Uniformization and Rao and Teh's algorithm

In this section we describe a classical and well-known technique of uniformization (Jensen, 1953).

A Markov jump process can be represented in terms of potential times of jumps and the corresponding states. Every trajectory $X([t^{\min}, t^{\max}])$ is right continuous and piecewise constant: $X(t) = S_{i-1}$ for $T_{i-1} \le t < T_i$, where random variables T_i are such that $t^{\min} < T_1 < \cdots < T_N < t^{\max}$ (by convention, $T_0 = t^{\min}$ and $t^{\max} < T_{N+1}$). The random sequence of states $S = (S_0, S_1, \ldots, S_N)$ such that $S_i = X(T_i)$ is called a skeleton. We do not assume that $S_{i-1} \neq S_i$, and therefore the two sequences

$$\begin{pmatrix} T \\ S \end{pmatrix} = \begin{pmatrix} t^{\min} & T_1 & \cdots & T_i & \cdots & T_N & t^{\max} \\ S_0 & S_1 & \cdots & S_i & \cdots & S_N & \end{pmatrix}$$

represent the process X in a redundant way: many pairs (T, S) correspond to the same trajectory $X([t^{\min}, t^{\max}])$. Let $J = \{i \in [1:N] : S_{i-1} \neq S_i\} \cup \{0\}$, so that $T_J = (T_i : i \in J)$ are moments of *true* jumps and $T_{-J} = T \setminus T_J = (T_i : i \notin J)$ are *virtual* jumps. We write $[l:r] = \{l, l+1, \ldots, r\}$. By a harmless abuse of notation, we identify increasing sequences of points in $[t^{\min}, t^{\max}]$ with finite sets. Note that the trajectory of X is uniquely defined by (T_J, S_J). Let us write $X \equiv (T_J, S_J)$ and also use the notation $J(X) = T_J$ for the set of true jump times.

Uniformization obtains if T is a sequence of consecutive points of a homogeneous Poisson process with intensity λ, where $\lambda \ge Q^{\max} = \max_s Q(s)$. The skeleton S is then (independently of T) a discrete-time, homogeneous Markov chain with the initial distribution ν and the transition matrix

$$P(s, s') = \begin{cases} \dfrac{Q(s, s')}{\lambda} & \text{if } s \neq s'; \\ 1 - \dfrac{Q(s)}{\lambda} & \text{if } s = s'. \end{cases} \tag{2}$$

Rao and Teh (2013) exploit uniformization to construct a special version of Gibbs sampler which converges to the posterior $p(X|Y)$. The key facts behind their algorithm are the following. First, given the trajectory $X \equiv (T_J, S_J)$ the conditional distribution of virtual jump times T_{-J} is that of the non-homogeneous (actually piecewise homogeneous) Poisson process with intensity $\lambda - Q(X(t)) \ge 0$. Second, this distribution does not change if we introduce the likelihood. Indeed, $L(Y|X) = L(Y|T_J, S_J)$, so Y and T_{-J} are conditionally independent given (T_J, S_J) and thus $p(T_{-J}|T_J, S_J, Y) = p(T_{-J}|T_J, S_J)$. Third, the conditional distribution $p(S|T, Y)$ is that of a hidden discrete time Markov chain and can be efficiently sampled from using the algorithm FFBS (Forward Filtering–Backward Sampling, Carter and Kohn (1994) and Frühwirth-Schnatter (1994)).

The Rao and Teh's algorithm generates a Markov chain $X_0, X_1, \ldots, X_m, \ldots$ (where $X_m = X_m([t^{\min}, t^{\max}])$ is a trajectory of an MJP), convergent to $p(X|Y)$, where $Y = (Y_1, \ldots, Y_k)$ is a vector of observations with the probability distribution of the form (1). A single step, that is the rule of transition from $X_{m-1} = X$ to $X_m = X'$ is described in Algorithm 1.

Convergence of the algorithm has been shown by its authors in Rao and Teh (2013). It follows from the fact that the chain has the stationary distribution $p(X|Y)$ and is irreducible and aperiodic, provided that $\lambda > Q^{\max}$.

Algorithm 1 Single step of Rao and Teh's algorithm.

input: previous state $(T_J, S_J) \equiv X$ and observation Y.
(V) Sample a Poisson process V with intensity $\lambda - Q(X(t))$ on $[t^{\min}, t^{\max}]$. Let $T' = T_J \cup V$ {*new set of potential times of jumps*}.
(S) Draw new skeleton S' from the conditional distribution $p(S'|T', Y)$ by FFBS. The new allocation of virtual and true jumps is via $J' = \{i : S'_{i-1} \neq S'_i\} \cup \{0\}$ {*we discard new virtual jumps* $T'_{-J'}$}.
return new state $(T'_{J'}, S'_{J'}) \equiv X'$.

4. Main result

Consider the Markov chain $X_0, X_1, \ldots, X_m, \ldots$ generated by the Rao and Teh's algorithm. Its transition kernel is denoted by A. Put differently, $A(X, \mathrm{d}X')$ is the probability distribution corresponding to Algorithm 1. Let $\Pi(\mathrm{d}X)$ be the posterior distribution of X given Y. In this paper we consider only Monte Carlo randomness, so Y is fixed and can be omitted in notation. The standing assumption is that $L(Y|X) > 0$ happens with nonzero probability if X is given by ν and Q. It means that the hidden MJP under consideration is "possible".

Theorem 1. *Assume that the matrix of intensities Q is irreducible and $\lambda > Q^{\max}$. Then the chain is geometrically ergodic, i.e. there exist constant $\gamma < 1$ and function M such that for every X,*

$$\|A^m(X, \cdot) - \Pi(\cdot)\|_{\mathrm{tv}} \leq \gamma^m M(X).$$

Let us begin with a brief outline of the main intuitions behind the proof. Note that we use uniformization with intensity λ of the dominating Poisson process strictly greater than $Q^{\max}$. Consequently, this process can be thought as a superposition of a Poisson process with intensity $\lambda^0 = \lambda - Q^{\max}$ and "the rest". Moreover, the transition matrix P has the diagonal elements bounded below by $\eta = 1 - Q^{\max}/\lambda > 0$. This has the following consequences.

On the one hand, on the set $\{X : |J(X)| \leq h\}$ there is a uniformly lower bounded probability of regeneration. Regeneration obtains if the following two events simultaneously occur. First, the new trajectory X' includes only jump times from the Poisson process with intensity λ^0, independent of X. Second, the old jump times disappear at stage (S), due to the fact that $\eta > 0$. This is formalized in Proposition 1.2.

On the other hand, it can be shown that a significant fraction (roughly proportional to η) of potential jump times becomes virtual and thus disappear. This gives a drift condition to the set $\{X : |J(X)| \leq h\}$ (Proposition 1.1).

Although the basic ideas are simple, the proofs are complicated because we have to take into account the restrictions imposed by the likelihood (i.e. consider the posterior probabilities). We need some auxiliary results.

Lemma 1.1. *Let $S_0, S_1, \ldots, S_n$ be a Markov chain on a finite state space $\mathcal{S}$, with transition matrix P. Assume that P is irreducible and $P(s, s) \geq \eta > 0$ for all $s \in \mathcal{S}$. Let $J = \{i \in [1 : n] : S_i \neq S_{i-1}\}$. There exist n_0 and $\delta > 0$ such that for every $n \geq n_0$ and for all $s_0, s_n \in \mathcal{S}$,*

$$\mathbb{E}(|J| \,|S_0 = s_0, S_n = s_n) \leq (1 - \delta)n.$$

Proof. Under the assumptions of the lemma, the Markov chain is ergodic (and also uniformly ergodic, because the state space is finite). Let π be its (strictly positive) stationary distribution. Fix $\varepsilon > 0$. There exists n_0 such that for $n \geq n_0/2 - 1$ we have for all $s_0, s_n \in \mathcal{S}$,

$$(1 - \varepsilon)\pi(s_n) \leq \mathbb{P}(S_n = s_n|S_0 = s_0) \leq (1 + \varepsilon)\pi(s_n).$$

Now if $n \geq n_0$ then we can choose n_1 such that $n_1 \geq n_0/2$ and $n - n_1 \geq n_0/2 - 1$. Let $J_1 = \{i \in [1 : n_1] : S_i \neq S_{i-1}\}$. With the notation $\rho(j, s) = \mathbb{P}(|J_1| = j, S_{n_1} = s|S_0 = s_0)$ we have

$$\begin{aligned}
\mathbb{E}(|J_1|\mathbb{I}(S_n = s_n)|S_0 = s_0) &= \sum_j \sum_s j\rho(j, s)P^{n-n_1}(s, s_n) \\
&\leq \sum_j \sum_s j\rho(j, s)(1 + \varepsilon)\pi(s_n) \\
&= \mathbb{E}(|J_1| \,|S_0 = s_0)(1 + \varepsilon)\pi(s_n) \\
&\leq (1 - \eta)n_1(1 + \varepsilon)\pi(s_n).
\end{aligned}$$

The last inequality follows from the fact that $\mathbb{P}(S_i = s|S_{i-1} = s) \geq \eta$. Consequently,

$$\begin{aligned}
\mathbb{E}(|J_1| \,|S_n = s_n, S_0 = s_0) &= \frac{\mathbb{E}(|J_1|\mathbb{I}(S_n = s_n)|S_0 = s_0)}{\mathbb{P}(S_n = s_n|S_0 = s_0)} \\
&\leq (1 - \eta)\frac{1 + \varepsilon}{1 - \varepsilon}n_1 = (1 - 2\delta)n_1,
\end{aligned}$$

where $\delta > 0$ if ε is chosen sufficiently small. Finally, since $|J| \le |J_1| + n - n_1$ and $n_1 \ge n/2$, we obtain

$$\mathbb{E}(|J| \,|S_n = s_n, S_0 = s_0) \le (1 - 2\delta)n_1 + n - n_1 = n - 2\delta n_1 \le (1 - \delta)n.$$

We get the conclusion.

Lemma 1.2. *Let the assumptions and the definitions of n_0 and δ in Lemma 1.1 hold. Let k be fixed. If $n \ge (k+1)n_0$, then for arbitrarily chosen indices $0 \le i_1 \le i_2 \le \cdots \le i_k \le n$ and states $s_1, \ldots, s_k$ we have that*

$$\mathbb{E}(|J| \,|S_{i_1} = s_1, \ldots, S_{i_k} = s_k) \le \left(1 - \frac{\delta}{k+1}\right) n,$$

provided that $\mathbb{P}(S_{i_1} = s_1, \ldots, S_{i_k} = s_k) > 0$.

Proof. Write $n_j = i_j - i_{j-1}$, where by convention $i_0 = 0$ and $i_{k+1} = n$. Let $J^{(j)} = \{i \in [i_{j-1} + 1 : i_j] : S_i \ne S_{i-1}\}$ for $j = 1, \ldots, k+1$. (Let us mention that we have not excluded the case when $i_{j-1} = i_j$ for some j and thus $J^{(j)} = \emptyset$. Of course, then we must have $s_{j-1} = s_j$ in the conditional expectation.) If $n \ge (k+1)n_0$ then there is at least one j such that $n_j \ge n/(k+1) \ge n_0$, because $\sum_j n_j = n$. From Lemma 1.1 we infer that for this j we have

$$\mathbb{E}(|J^{(j)}| \,|S_{i_1} = s_1, \ldots, S_{i_k} = s_k) = \mathbb{E}(|J^{(j)}| \,|S_{i_{j-1}} = s_{j-1}, S_{i_j} = s_j) \le (1 - \delta)n_j.$$

The equality above is true because of the (two-sided) Markov property. Analogously as in the previous proof, we observe that $|J| \le |J^{(j)}| + n - n_j$, so

$$\mathbb{E}(|J| \,|S_{i_1} = s_1, \ldots, S_{i_k} = s_k) \le (1 - \delta)n_j + n - n_j = n - \delta n_j \le \left(1 - \frac{\delta}{k+1}\right) n.$$

In the next proposition we establish a geometric drift condition for the Markov chain $X_0, X_1, \ldots, X_m, \ldots$. Consider a single step, that is transition from $X_{m-1} = X$ to $X_m = X'$. Thus only Monte Carlo randomness in Algorithm 1 is taken into account. The dependence on the input trajectory X (and on Y) is implicitly assumed but indicated only when necessary. Recall that $|J(X)|$ is the number of true jumps of the trajectory $X([t^{\min}, t^{\max}])$.

Proposition 1.1 (*Drift Condition*). *There exist $q < 1$ and $c < \infty$ such that in a single step of the Rao and Teh's algorithm we have $\mathbb{E}(|J(X')| \,|X) \le q|J(X)| + c$.*

Proof. Let us analyse what happens in two stages of Algorithm 1. We fix the initial X. In the first stage we add a new set V of potential jumps. Since $|V|$ has the Poisson distribution with intensity $\int_{t^{\min}}^{t^{\max}} (\lambda - Q(X(t)))\mathrm{d}t$, we have $\mathbb{E}|V| \le \lambda(t^{\max} - t^{\min}) := \mu$. Thus we obtain T' with $\mathbb{E}|T'| \le |J(X)| + \mu$. In the second stage T' is "thinned" to $T'_{J'}$. We will prove that

$$\mathbb{E}(|J'| \,|T') \le \left(1 - \frac{\delta}{k+1}\right) |T'|, \quad \text{if } |T'| \ge (k+1)n_0. \tag{3}$$

Inequality (3) implies $\mathbb{E}(|J(X')|) = \mathbb{E}(|J'|) \le (1 - \delta/(k+1))(|J(X)| + \mu)$. The conclusion of the theorem follows with $q = (1 - \delta/(k+1))$. To ensure that the conclusion holds also if $|T'| < (k+1)n_0$, we can choose $c = q\mu + (k+1)n_0$ so that $|J'| < c$.

It remains to show (3). We consider the second stage of Algorithm 1 (sampling a new skeleton S' from $p(S'|T', Y)$). From now on, the result of the first stage (updating T to T') is fixed. Note that our assumption about the structure of observations (1) implies that $p(S'|T', Y) \propto p(S'|T') \prod_{j=1}^{k} L_j(Y_j|S'_{i_j})$, where $i_j = \max\{i : T'_i \le t_j^{\text{obs}}\}$.

Since $p(S'|T')$ is the distribution of a Markov chain, it follows that

$$p(S'|T', Y) \propto \nu(S'_0) \prod_{i=1}^{|T'|} P(S'_{i-1}, S'_i) \prod_{j=1}^{k} L_j(S'_{i_j}),$$

where P is given by (2) and $L_j(s) = L_j(Y_j|s)$.

Although the actual sampling from $p(S'|T', Y)$ is by FFBS, exactly the same result can be obtained via rejection sampling as follows.

(S1) Simulate Markov chain S' (of length $|T'|$) with transition matrix P given by (2) and initial distribution ν.

(S2) The skeleton S' is accepted with probability $\prod_{j=1}^{k}(L_j(S'_{i_j})/L_j^{\max})$, with some $L_j^{\max} \ge \max_s L_j(s)$. Otherwise go to (S1).

(Of course the rejection method is highly inefficient and is considered only to clarify presentation.) Now we are in a position to use Lemma 1.2, with $\eta = 1 - Q^{\max}/\lambda$. In the formulas to follow, we implicitly condition on the event $|T'| = n$ without

indicating this in notation. If $n \geq (k+1)n_0$ then

$$\begin{aligned}\mathbb{E}(|J'|) &= \mathbb{E}(|J'| \,|S' \text{accepted}) \\ &= \sum_{s_1,\ldots,s_k} \mathbb{E}(|J'| \,|S'_{i_1} = s_1, \ldots, S'_{i_k} = s_k)\mathbb{P}(S'_{i_1} = s_1, \ldots, S'_{i_k} = s_k|S' \text{accepted}) \\ &\leq \left(1 - \frac{\delta}{k+1}\right) n \sum_{s_1,\ldots,s_k} \mathbb{P}(S'_{i_1} = s_1, \ldots, S'_{i_k} = s_k|S' \text{accepted}) \\ &\leq \left(1 - \frac{\delta}{k+1}\right) n\end{aligned}$$

(of course in the sum above we can omit "impossible sequences" $s_1, \ldots, s_k$). We have proved (3) and we are done.

Remark 1.1. The proof of Proposition 1.1 could be significantly simpler if we assumed that the likelihood $L_j(s)$ is strictly positive. However, this assumption is not satisfied in many interesting applications (e.g. if the observations are without noise).

Remark 1.2. The efficiency of Rao and Teh's algorithm strongly depends on the choice of λ. Let us comment on the role λ plays in our proofs. The larger λ the larger $\eta = 1 - Q^{\max}/\lambda$. Increasing η gives better estimates in Lemma 1.2 and consequently in Proposition 1.1. On the other hand, larger λ leads to slower mixing rate for the skeleton chain, i.e. increases n_0 in Lemma 1.2 and thus worsens the estimates.

Proposition 1.2 (*Minorization Condition*)**.** *The set $\{X : |J(X)| \leq h\}$ is 1-small for every h, i.e. there exists a probability measure R and a constant $\beta > 0$ such that $A(X, \mathrm{d}X') \geq \beta R(\mathrm{d}X')$, whenever $|J(X)| \leq h$.*

Recall that A denotes the transition kernel of the Markov chain defined via Algorithm 1. R is a called a regeneration measure.

Proof. The schema of our proof is the following. We will define a sequence of states $s^* = (s_0^*, \ldots, s_l^*)$ and a sequence of times $t^* = (t_0^*, \ldots, t_l^*)$. Both these sequences are deterministic and fixed. The regeneration measure $R(\mathrm{d}X')$ is described in terms of s^* and t^* as follows:

$$\begin{aligned}&T'_i \sim \text{Uniform}(t_{i-1}^*, t_i^*) \text{independently} \quad \text{for } i = 1, \ldots, l; \\ &S'_i = s_i^* \quad \text{for } i = 0, 1, \ldots, l.\end{aligned} \tag{4}$$

Trajectory X' is determined by (T', S') as described in Section 2. Note that the skeleton S' is deterministic and random vector $(T'_1, \ldots, T'_l)$ has the uniform distribution on the set

$$\mathcal{T} = \{(t_1, \ldots, t_l) : t_{i-1}^* \leq t_i \leq t_i^* \text{ for } i = 1, \ldots, l\}.$$

We will show that Algorithm 1 can be equivalently executed in such a way that the resulting X' is distributed according to R with probability β, provided that $|J(X)| \leq h$ (β must not depend on X; it will be defined in the course of our proof).

Now we proceed to details of our construction. To define s^* and t^*, let us first choose a sequence $s^\dagger = (s_1^\dagger, s_2^\dagger, \ldots, s_k^\dagger)$ such that

$$\prod_{j=1}^{k} (L_j(s_j^\dagger)/L_j^{\max}) = \beta_1 > 0.$$

By irreducibility of kernel P we can embed $s^\dagger$ in a possible skeleton s^*, i.e. we define a sequence $s^* = (s_0^*, \ldots, s_l^*)$ for some $l \geq k$ such that $s^\dagger$ is a subsequence of s^*, $s_{i-1}^* \neq s_i^*$ and

$$\nu(s_0^*) \prod_{i=1}^{l} P(s_i^*, s_{i+1}^*) = \beta_2 > 0.$$

Similarly we embed the sequence $t^{\text{obs}} = (t_1^{\text{obs}}, \ldots, t_k^{\text{obs}})$ in a longer sequence $t^* = (t_0^*, t_1^*, \ldots, t_l^*)$. More precisely, we choose a sequence $t^{\min} = t_0^* < t_1^* < \cdots < t_l^* < t^{\max}$ such that $s_{i_j}^* = s_j^\dagger$ implies $t_{i_j}^* = t_j^{\text{obs}}$ for $j = 1, \ldots, k$.

Fix X with $|J(X)| \leq h$. We are going to describe a special way in which Algorithm 1 can be executed. In stage (V) we can independently sample two Poisson processes on the interval $[t^{\min}, t^{\max}]$, say V^0 and V^{rest}, with intensities $\lambda - Q^{\max}$ and $Q^{\max} - Q(X(t))$, respectively. Next let $V = V^0 \cup V^{\text{rest}}$ and $T' = J(X) \cup V$. Note that

$$\mathbb{P}(V^0 \in \mathcal{T}) = \beta_3 > 0.$$

In stage (S) of Algorithm 1 we construct skeleton S'. Assume that we use rejection sampling, defined by (S1) and (S2) in the previous proof. We use the notation as in Algorithm 1. Recall that $J' = \{i : S'_{i-1} \neq S'_i\}$. We will bound from below the

probability that at stage (S1) all points belonging to $V^{\text{rest}} \cup J(X)$ are changed to virtual jumps, while jumps at V^0 form the skeleton s^*. We have

$$\mathbb{P}(T'_{J'} = V^0, S'_{J'} = s^*|T') \geq \beta_2 \sum_{k=0}^{\infty} \eta^{|J(X)|+k} \mathbb{P}(V^{\text{rest}} = k|X)$$
$$\geq \beta_2 \mathbb{E}(\eta^{h+|V^{\text{rest}}|}|X) \geq \beta_2 \eta^{h+\mathbb{E}(|V^{\text{rest}}|\,|X)}$$

by Jensen inequality. Since V^{rest} is a Poisson process with intensity bounded by $Q^{\max}$ we have $\mathbb{E}(|V^{\text{rest}}|\,|X) \leq Q^{\max}(t^{\max} - t^{\min})$, therefore we conclude that $\mathbb{P}(T'_{J'} = V^0, S'_{J'} = s^*|T') \geq \beta_4 > 0$. The probability of accepting the obtained skeleton $S' = s^*$ at stage (S2) is clearly equal to β_1. We have shown that $\mathbb{P}(V^0 \in \mathcal{T}, T'_{J'} = V^0, S'_{J'} = s^*|X) \geq \beta_1\beta_4\beta_3 = \beta > 0$ whenever $|J(X)| \leq h$. Moreover, if $T'_{J'} = V^0$ and $S'_{J'} = s^*$ then the probability distribution of $(T'_{J'}, S'_{J'}) \equiv X'$ does not depend on X and is equal to R. This completes the proof.

Remark 1.3. The best choice of λ is a delicate problem. It can be seen from our proof of Proposition 1.2 that β decreases if $\lambda \searrow Q^{\max}$ and also if $\lambda \nearrow \infty$. The constants q and c in the drift condition are also influenced by λ, see Remark 1.1. Finally, the cost of a single step of the algorithm increases linearly with λ.

Proof of Theorem 1. Theorem 1 follows from Propositions 1.1 and 1.2. We will use Roberts and Rosenthal (2004, Th. 9). We only need to verify a slightly different but qualitatively equivalent version of drift condition, which appears in the cited paper. From Proposition 1.1 it follows that

$$\mathbb{E}(|J(X')|\,|X) \leq q'|J(X)| + c\mathbb{I}(|J(X)| > h),$$

if $1 > q' \geq q + c/h$. This can be ensured if we choose $h > c/(1-q)$ in the definition of a small set.

Acknowledgements

The authors are grateful to the anonymous referees whose suggestions helped improve the paper. The work of Błażej Miasojedow is supported by Polish National Science Center Grant No. 2015/17/D/ST1/01198.

References

Boys, R.J., Wilkinson, D.J., Kirkwood, T.B., 2008. Bayesian inference for a discretely observed stochastic kinetic model. Stat. Comput. 18, 125–135.
Carter, C., Kohn, R., 1994. On Gibbs sampling for state space models. Biometrika 81, 541–553. URL: http://www.jstor.org/stable/2337125.
El-Hay, T., Friedman, N., Kupferman, R., 2008. Gibbs sampling in factorized continuous-time Markov processes. In: Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence. UAI-08. AUAI Press, Corvallis, Oregon, pp. 169–178.
Fan, Y., Shelton, C.R., Sampling for approximate inference in continuous time Bayesian networks, in: Tenth International Symposium on Artificial Intelligence and Mathematics, 2008.
Fearnhead, P., Sherlock, C., 2006. An exact Gibbs sampler for the Markov-modulated Poisson process. J. R. Stat. Soc. Ser. B Stat. Methodol. 68, 767–784. http://dx.doi.org/10.1111/j.1467-9868.2006.00566.x.
Frühwirth-Schnatter, S., 1994. Data augmentation and dynamic linear models. J. Time Series Anal. 15, 183–202. http://dx.doi.org/10.1111/j.1467-9892.1994.tb00184.x.
Golightly, A., Henderson, D., Sherlock, C., 2015. Delayed acceptance particle MCMC for exact inference in stochastic kinetic models. Stat. Comput. 25, 1039–1055. http://dx.doi.org/10.1007/s11222-014-9469-x.
Golightly, A., Wilkinson, D.J., 2011. Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. Interface Focus http://dx.doi.org/10.1098/rsfs.2011.0047.
Golightly, A., Wilkinson, D.J., Bayesian inference for Markov jump processes with informative observations, 2014. ArXiv e-prints.
Jensen, A., 1953. Markoff chains as an aid in the study of Markoff processes. Scand. Actuar. J. 1953, 87–91.
Nodelman, U., Shelton, C.R., Koller, D., 2002. Learning continuous time Bayesian networks. In: Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., pp. 451–458.
Rao, V., Teh, Y.W., 2012. MCMC for continuous-time discrete-state systems. In: Advances in Neural Information Processing Systems. pp. 701–709.
Rao, V., Teh, Y.W., 2013. Fast MCMC sampling for Markov jump processes and extensions. J. Mach. Learn. Res. 14, 3207–3232.
Roberts, G.O., Rosenthal, J.S., 2004. General state space Markov chains and MCMC algorithms. Probab. Surv. 1, 20–71.