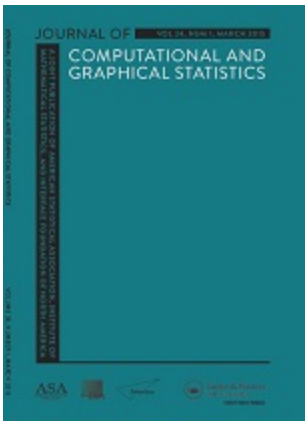

Efficient Parameter Sampling for Markov Jump Processes

Journal:	*Journal of Computational and Graphical Statistics*
Manuscript ID	JCGS-18-220.R2
Manuscript Type:	Original Article
Keywords:	Continuous-time Markov chain, Markov Chain Monte Carlo (MCMC), Metropolis Hastings, Uniformization, Geometric ergodocity

SCHOLARONE™
Manuscripts

Efficient Parameter Sampling for Markov Jump Processes

Boqian Zhang and Vinayak Rao,
Department of Statistics, Purdue University, USA
email: zhan1977@purdue.edu, varao@purdue.edu

March 25, 2020

For Peer Review Only

Abstract

Markov jump processes are continuous-time stochastic processes widely used in a variety of applied disciplines. Inference typically proceeds via Markov chain Monte Carlo, the state-of-the-art being a uniformization-based auxiliary variable Gibbs sampler. This was designed for situations where the process parameters are known, and Bayesian inference over unknown parameters is typically carried out by incorporating it into a larger Gibbs sampler. This strategy of sampling parameters given path, and path given parameters can result in poor Markov chain mixing. In this work, we propose a simple and efficient algorithm to address this problem. Our scheme brings Metropolis-Hastings approaches for discrete-time hidden Markov models to the continuous-time setting, resulting in a complete and clean recipe for parameter and path inference in Markov jump processes. In our experiments, we demonstrate superior performance over Gibbs sampling, a more naïve Metropolis-Hastings algorithm, as well as another popular approach, particle Markov chain Monte Carlo. We also show our sampler inherits geometric mixing from an 'ideal' sampler that is computationally much more expensive. Supplementary material for the article is available online.

Keywords: Continuous-time Markov chain, Markov chain Monte Carlo, Metropolis-Hastings, Uniformization, Geometric Ergodicity

1 Introduction

Markov jump processes (MJPs) are continuous-time stochastic processes widely used in fields like computational chemistry (Gillespie, 1977), molecular genetics (Fearnhead and Sherlock, 2006), mathematical finance (Elliott and Osakwe, 2006), queuing theory (Breuer, 2003), artificial intelligence (Xu and Shelton, 2010) and social-network analysis (Pan et al., 2016). MJPs have been used as realistic, mechanistic and interpretable models of a wide variety of phenomena, among others, the references above have used them to model temporal evolution of the state of a chemical reaction or queuing network, segmentation of a strand of DNA, and user activity on social media. Their continuous-time dynamics however raise computational challenges when, given noisy measurements, one wants to make inferences over the latent MJP trajectory as well as any process parameters. In contrast to discrete-time hidden Markov models, one cannot *a priori* bound the number of trajectory state transitions, and the transition times themselves are continuous-valued. The state-of-the-art approach is an auxiliary variable Gibbs sampler from Rao and Teh (2013), we will refer to this as the Rao-Teh algorithm. This Markov chain Monte Carlo (MCMC) algorithm was designed to simulate paths when the MJP parameters are known. Parameter inference is typically carried out by incorporating it into a Gibbs sampler that also conditionally simulates parameters given the currently sampled trajectory.

In many situations, the MJP trajectory and parameters exhibit strong coupling, so that alternately sampling path given parameters, and parameters given path can result in poor mixing. To address this, we propose an efficient Metropolis-Hastings (MH) sampler (algorithm 4). In our experiments, we demonstrate superior performance over Gibbs sampling, a more naïve MH sampler (algorithm 3), as well as particle Markov chain Monte Carlo (Andrieu et al., 2010). We also prove that under relatively mild conditions, our sampler inherits geometric ergodicity from an 'ideal' sampler that is computationally much more expensive.

For Peer Review Only

2 Markov jump processes (MJPs)

A Markov jump process (Çinlar, 1975) is a right-continuous piecewise-constant stochastic process taking values in a state space $\mathcal{S}$. We assume a finite number of states N, with $\mathcal{S} = \{1, \ldots, N\}$. Then, the MJP is parameterized by two quantities, an N-component probability vector π_0 and a rate-matrix A. The former gives the distribution over states at the initial time (we assume this is 0), while the latter is an $N \times N$-matrix governing the dynamics of the system. An off-diagonal element A_{ij} gives the rate of transitioning from state i to j. The rows of A sum to 0, so that $A_{ii} = -\sum_{j \neq i} A_{ij}$. We write A_i for the negative of the ith diagonal element A_{ii}, so that $A_i = -A_{ii}$ gives the total rate at which the system leaves state i for any other state. To simulate an MJP over an interval $[0, t_{end})$, one follows Gillespie's algorithm (Gillespie, 1977): first sample an initial state s_0 from π_0, and defining $t_0 = t_{curr} = 0$ and $k = 0$, repeat the following while $t_{curr} < t_{end}$:

- Simulate a wait-time Δt_k from an exponential distribution with rate A_{s_k}. Set $t_{k+1} = t_{curr} = t_k + \Delta t_k$. The MJP remains in state s_k until time t_{k+1}.
- Jump to a new state $s_{k+1} \neq s_k$ with probability equal to $A_{s_k s_{k+1}}/A_{s_k}$. Set $k = k + 1$.

The times $T = (t_1, \ldots, t_{k-1})$ and states $S = (s_1, \ldots, s_{k-1})$, along with the initial state s_0, define the MJP path. We use both (s_0, S, T) and $\{S(t), t \in [0, t_{end})\}$ (and sometimes just $S(\cdot)$) to refer to the MJP path. See the top-left panel in figure 2 for a sample path.

2.1 Structured rate matrices

While the rate matrix A can have $N(N-1)$ independent elements, in typical applications, especially with large state-spaces, it is determined by a much smaller set of parameters. We will write these as θ, with A a deterministic function of these parameters: $A \equiv A(\theta)$. The parameters θ are often more interpretable than the elements of A, and correspond directly to physical, biological or environmental parameters of interest. For example:

Immigration-death processes Here, $\theta = (\alpha, \beta)$, with α the arrival-rate and β the death-rate. The state represents the size of a population or queue. New individuals enter

For Peer Review Only

with rate α, so off-diagonal elements $A_{i,i+1}$ equal α. Each individual dies at a rate β, so that $A_{i,i-1} = i\beta$ for each i. All other transitions have rate 0.

Birth-death processes This variant of the earlier MJP moves from state i to $i+1$ with rate $i\alpha$, with growth-rate proportional to population size. The death-rate is β, so that $A_{i,i-1} = i\beta$ for each i. Other off-diagonal elements are 0, and again $\theta = (\alpha, \beta)$.

Codon substitution models These characterize transitions between codons at a DNA locus over evolutionary time. There are 61 codons, and in the simplest case, all transitions have the same rate (Jukes and Cantor, 1969): $A_{ij} = \alpha \; \forall i \neq j$. Thus the 61×61 matrix A is determined by a single α. Other models (Goldman and Yang, 1994) group transitions as 'synonymous' and 'nonsynonymous', based on whether old and new codons encode the same amino acid. Synonymous and nonsynonymous transitions have their own rates, so A is now determined by 2 parameters α and β.

3 Bayesian modeling and inference for MJPs

We first set up our Bayesian model of the data generation process. We model a latent piecewise-constant path $S(\cdot)$ over $[0, t_{end})$ as an N-state MJP with rate matrix $A(\theta)$ and prior π_0 over $s_0 = S(0)$, the state at time 0. We place a prior $P(\theta)$ over the unknown θ. For simplicity, we assume π_0 is known (or we set it to a uniform distribution over the N states). We have noisy measurements X of the latent process, with likelihood $P(X|\{S(t),\ t \in [0, t_{end})\})$. Again, for clarity we ignore any unknown parameters in the likelihood, else we can include them in θ. We assume the observation process has the following structure: for fixed X, for any partition $\tilde{W} = \{\tilde{w}_1 = 0, \ldots, \tilde{w}_{|\tilde{W}|} = t_{end}\}$ of the interval $[0, t_{end})$ (where $|\cdot|$ denotes cardinality), there exist known functions ℓ_i such that the likelihood factors as:

$$P(X|\{S(t),\ t \in [0, t_{end})\}) = \prod_{i=1}^{|\tilde{W}|-1} \ell_i(\{S(t),\ t \in [\tilde{w}_i, \tilde{w}_{i+1})\}) \tag{1}$$

A common example is a finite set of independent observations $X = \{x_1, \ldots, x_{|X|}\}$ at times $T^X = \{t_1^X, \ldots, t_{|X|}^X\}$, each observation depending on the state of the MJP at that time:

For Peer Review Only

$$P(X|\{S(t),\ t \in [0, t_{end})\}) = \prod_{i=1}^{|X|} P(x_i|S(t_i^X)). \tag{2}$$

Other examples are an inhomogeneous Poisson process (Fearnhead and Sherlock, 2006), renewal process (Rao and Teh, 2011) or even another MJP (Nodelman et al., 2002; Rao and Teh, 2013), modulated by (s_0, S, T). The first example, called a Markov modulated Poisson process (MMPP) (Scott and Smyth, 2003), associates a positive rate λ_s with each state s, with $\ell_i(\{S(t),\ t \in [w_i, w_{i+1})\})$ equal to the likelihood of the Poisson events within $[w_i, w_{i+1})$ under an inhomogeneous Poisson process with piecewise-constant rate $\{\lambda_{S(t)},\ t \in [w_i, w_{i+1})\}$.

With $A(\cdot)$ and π_0 assumed known, the overall Bayesian model is then

$$\theta \sim P(\theta), \quad (s_0, S, T) \sim \mathrm{MJP}(\pi_0, A(\theta)), \quad X \sim P(X|s_0, S, T). \tag{3}$$

Given X, one is interested in the posterior distribution over the latent quantities, (θ, s_0, S, T).

3.1 Trajectory inference given the MJP parameters θ

This was addressed in Rao and Teh (2013) and extended to a broader class of jump processes in Rao and Teh (2012) (also see Fearnhead and Sherlock, 2006; Hobolth and Stone, 2009; El-Hay et al., 2008)). Rao and Teh (2013, 2012) both involve MJP path representations with auxiliary *candidate* jump times that are later *thinned*. We focus on the former, a simpler and more popular algorithm, based on the idea of *uniformization* (Jensen, 1953).

Uniformization involves a parameter $\Omega(\theta) \geq \max_i A_i(\theta)$; Rao and Teh (2013) suggest $\Omega(\theta) = 2\max_i A_i(\theta)$. Define $B(\theta) = \left(I + \frac{1}{\Omega(\theta)} A(\theta)\right)$; this is a stochastic matrix with nonnegative elements, and rows adding up to 1. Unlike the sequential wait-and-jump Gillespie algorithm, uniformization first simulates a random grid of candidate transition-times W over $[0, t_{end})$, and then assigns these state values:

- Simulate W from a Poisson process with rate $\Omega(\theta) \geq \max_i A_i(\theta)$: $W \sim \mathrm{PoissProc}(\Omega(\theta))$.
- Assign states (v_0, V) to the times $0 \cup W$, with $v_0 \sim \pi_0$, and $P(v_{i+1} = s|v_i) = B_{v_i s}(\theta)$.

Setting $\Omega(\theta) > \max_i A_i(\theta)$ results in more candidate-times than actual MJP transitions, at the same time, unlike $A(\theta)$, the matrix $B(\theta)$ can thin these through self-transitions.

For Peer Review Only

Write U for the elements W with self-transitions, and T for the rest. Define $s_0 = v_0$, and $S = \{v_i \in V \text{ s.t. } v_i \neq v_{i-1}\}$ as the elements in V corresponding to T, then (s_0, S, T) sampled this way for any $\Omega(\theta) \geq \max_i A_i(\theta)$ has the same distribution as under Gillespie's algorithm (Jensen, 1953; Rao and Teh, 2013). The third panel in figure 2 shows these sets.

Introducing the thinned variables allowed Rao and Teh (2013) to develop an efficient MCMC sampler (algorithm 1). At a high-level, each MCMC iteration simulates a new grid W conditioned on the path (s_0, S, T), and then a new path conditioned on W. Rao and Teh (2013) show that the resulting Markov chain targets the desired posterior distribution over trajectories, and is ergodic for any $\Omega(\theta)$ strictly greater than all the $A_i(\theta)$'s.

Algorithm 1 The Rao and Teh (2013) MCMC sampler for MJP trajectories

Input: Prior π_0, observations X, the previous path (s_0, S, T).
Parameter $\Omega(\theta) > \max_i A_i(\theta)$, where $A(\theta)$ is the MJP rate-matrix.

Output: New MJP trajectory (s_0', S', T').

1: **Simulate the thinned candidate times U given the MJP path** (s_0, S, T) from a piecewise-constant rate-$(\Omega(\theta) - A_{S(\cdot)}(\theta))$ Poisson process:

$$U \sim \text{PoissProc}(\Omega(\theta) - A_{S(t)}(\theta)), \quad t \in [0, t_{end}).$$

2: **Discard the states (s_0, S), and write $W = T \cup U$.**

3: **Simulate states (v_0, V) on $0 \cup W$ from a discrete-time HMM** with initial distribution over states π_0 and transition matrix $B(\theta) = \left(I + \frac{1}{\Omega(\theta)} A(\theta)\right)$. Following equation (1), between two consecutive times $(\tilde{w}_i, \tilde{w}_{i+1})$ in $\tilde{W} \stackrel{\text{def}}{=} 0 \cup W \cup t_{end}$, state s has likelihood $\ell_i(s) \equiv \ell_i(\{S(t) = s,\ t \in [\tilde{w}_i, \tilde{w}_{i+1})\})$. The simulation involves two steps:

Forward pass: Set $\mathtt{f}_0(\cdot) = \pi_0$. Sequentially update $\mathtt{f}_i(\cdot)$ at time $w_i \in W$ given $\mathtt{f}_{i-1}$:

$$\textbf{for } i = 1 \to |W| \textbf{ do:} \quad \mathtt{f}_i(s') = \sum_{s \in \mathcal{S}} \mathtt{f}_{i-1}(s) \cdot \ell_i(s) \cdot B_{ss'}(\theta), \quad \forall s' \in \mathcal{S}.$$

Backward pass: Simulate $v_{|W|} \sim \mathtt{b}_{|W|}(\cdot)$, where $\mathtt{b}_{|W|}(s) \propto \mathtt{f}_{|W|}(s) \cdot \ell_{|W|+1}(s) \quad \forall s \in \mathcal{S}$.

$$\textbf{for } i = (|W|-1) \to 0 \textbf{ do:} \quad v_i \sim \mathtt{b}_i(\cdot), \text{ where } \mathtt{b}_i(s) \propto \mathtt{f}_i(s) \cdot B_{s v_{i+1}}(\theta) \cdot \ell_{i+1}(s) \quad \forall s \in \mathcal{S}.$$

4: **Discard self-transitions**: Set $s_0' = v_0$. Let T' be the set of times in W when V changes state. Define S' as the corresponding set of state values. Return (s_0', S', T').

For Peer Review Only

3.2 Joint inference over MJP path (s_0, S, T) and parameters θ

For fixed parameters θ, the efficiency of the Rao-Teh algorithm has been established, both empirically (Rao and Teh, 2013) and theoretically (Miasojedow and Niemiro, 2017). In practice, the parameters are typically unknown, and often, these are of primary interest. One then has to characterize the complete posterior $P(\theta, s_0, S, T|X)$ of the Bayesian model of equation (3). This is typically carried out by incorporating the previous algorithm into a Gibbs sampler that targets the joint $P(\theta, s_0, S, T|X)$ by conditionally simulating (s_0, S, T) given θ and then θ given (s_0, S, T). Algorithm 2 (see also Rao and Teh, 2013) outlines this:

Algorithm 2 Gibbs sampling for path and parameter inference for MJPs

Input: The current MJP path (s_0, S, T), the current MJP parameters θ.

Output: New MJP trajectory (s_0', S', T') and parameters θ'.

1: Simulate a new path from the conditional $P(s_0', S', T'|X, s_0, S, T, \theta)$ by algorithm 1.

2: Simulate a new parameter θ' from the conditional $P(\theta'|X, s_0', S', T')$ (see equation (4)).

The distribution $P(\theta'|X, s_0', S', T')$ depends on the amount of time τ_i spent in each state i, and the number of transitions c_{ij} between each pair of states i, j:

$$P(\theta'|X, s_0', S', T') \propto P(\theta') \prod_{i \in \mathcal{S}} \exp(-A_i(\theta')\tau_i) \prod_{j \in \mathcal{S}} \left(\frac{A_{ij}(\theta')}{A_i(\theta')} \right)^{c_{ij}}. \tag{4}$$

In some circumstances, this can be directly sampled from, otherwise, one has to use a Markov kernel like Metropolis-Hastings to update θ to θ'. In any event, this introduces no new technical challenges. However, the resulting Gibbs sampler can mix very poorly because of coupling between path and parameters. We illustrate this in figure 1 (inspired by Papaspiliopoulos et al., 2007), which shows the posterior distribution of an MJP parameter (long-dashes) is less concentrated than the distribution conditioned on both observations as well as path (short-dashes). The coupling is strengthened as the trajectory grows longer (right panel), and the Gibbs sampler can mix very poorly with long observation periods, even if the observations themselves are only mildly informative about the parameters. Before we describe our actual algorithm, we outline a naïve attempt around this coupling.

For Peer Review Only

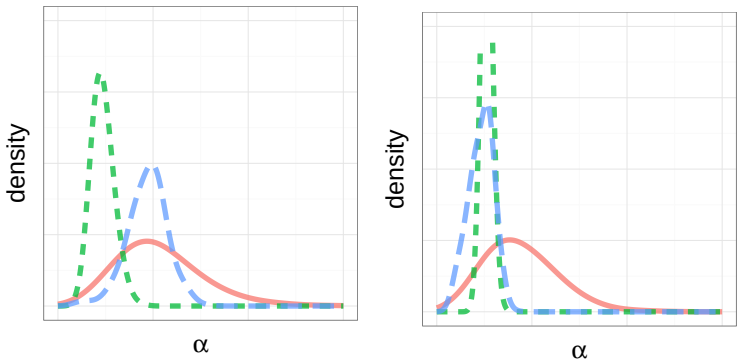

Figure 1: Prior density over an MJP parameter (solid curve), along with two conditionals: given observations only (long-dashes), and given observations and MJP path (short-dashes). As t_{end} increases from 10 (left) to 100 (right), the conditionals become more concentrated, implying stronger path-parameter coupling. The plots are from section 7.3 with 3 states.

4 Naïve parameter inference via Metropolis-Hastings

For discrete-time HMMs, path-parameter coupling can be circumvented by marginalizing out the Markov trajectory, and directly sampling from the marginal posterior $P(\theta|X)$. In its simplest form, this involves a Metropolis-Hastings (MH) scheme that proposes a new parameter ϑ from a proposal distribution $q(\vartheta|\theta)$, accepting or rejecting according to the usual MH probability. The marginal probabilities over X given parameters are computed using the forward-filtering backward-sampling (FFBS) algorithm (Frühwirth-Schnatter, 1994; Carter and Kohn, 1996; Rao and Teh, 2013). The Rao-Teh algorithm, which recasts posterior simulation for continuous-time models as discrete-time simulation on a random grid, then provides a simple way to incorporate such an MH-scheme into continuous-time settings: directly update θ, conditioning on the random grid W, but marginalizing out (v_0, V).

Specifically, given θ and the Poisson grid W, rather than simulating new path values (the backward pass in algorithm 1), and then conditionally updating θ (the second step in algorithm 2), we *first* propose a parameter ϑ from $q(\vartheta|\theta)$. This is accepted with probability

$$\mathtt{acc} = \min\left(1, \frac{P(X|W,\vartheta)P(W|\vartheta)P(\vartheta)q(\theta|\vartheta)}{P(X|W,\theta)P(W|\theta)P(\theta)q(\vartheta|\theta)}\right),$$

thereby targeting the distribution $P(\theta|W,X)$. In the equation above, $P(X|W,\theta)$ is the probability of the observations X given W with (v_0, V) marginalized out. Uniformization

For Peer Review Only

says this is the marginal probability of X under a discrete-time HMM on W, with transition matrix $B(\theta)$. This can be computed using the forward pass of FFBS algorithm (steps 4 and 6 of algorithm 3 below). The term $P(W|\theta)$ is the probability of W under a rate-$\Omega(\theta)$ Poisson process. These, and the corresponding terms for ϑ allow the acceptance probability to be computed. Only *after* accepting or rejecting ϑ do we simulate new states (v_0', V'), using the new parameter θ' in a backward pass over W. The new trajectory and parameter are used to simulate a new grid W', and the process is repeated. Algorithm 3 includes all details of this algorithm (see also figure 19 in the supplementary material).

Algorithm 3 Naïve MH for parameter inference for MJPs

Input: Observations X, the MJP path (s_0, S, T), the MJP parameters θ and π_0.

Output: A new MJP trajectory (s_0', S', T'), new MJP parameter θ'.

1: Set $\Omega(\theta) > \max_s A_s(\theta)$ for some function $\Omega(\cdot)$, e.g. $\Omega(\theta) = 2\max_s A_s(\theta)$.

2: **Simulate the thinned times** U from a rate-$(\Omega(\theta) - A_{S(\cdot)}(\theta))$ Poisson process:

$$U \sim \text{PoissProc}(\Omega(\theta) - A_{S(t)}(\theta)), \quad t \in [0, t_{end}).$$

3: Set $W = T \cup U$ and discard (s_0, S). Define $\tilde{W} = 0 \cup W \cup t_{end}$.

4: **Forward pass:** Set $B(\theta) = I + \frac{1}{\Omega(\theta)} A(\theta)$ and $\mathtt{f}_0^\theta(\cdot) = \pi_0$. Recall $\ell_i(\cdot)$ from equation (1).

for $i = 1 \to |W|$ **do:** $\mathtt{f}_i^\theta(s') = \sum_{s \in \mathcal{S}} \mathtt{f}_{i-1}^\theta(s) \cdot \ell_i(s) \cdot B_{ss'}(\theta), \quad \forall s' \in \mathcal{S}.$

5: **Propose** $\vartheta \sim q(\cdot|\theta)$**.** For elements of W, calculate $\mathtt{f}_i^\vartheta(\cdot)$ similar to above.

6: **Accept/Reject:** For θ (and similarly for ϑ), set $P(W|\theta) = \Omega(\theta)^{|W|} \exp(-\Omega(\theta) t_{end})$, $P(X|W, \theta) = \sum_{s \in \mathcal{S}} \mathtt{f}_{|W|}^\theta(s) \cdot \ell_{|W|+1}(s)$. With probability $\mathtt{acc}$, set $\theta' = \vartheta$, else $\theta' = \theta$;

$$\mathtt{acc} = 1 \wedge \frac{P(\vartheta|W, X)}{P(\theta|W, X)} \frac{q(\theta|\vartheta)}{q(\vartheta|\theta)} = 1 \wedge \frac{P(X|W, \vartheta) P(W|\vartheta) P(\vartheta)}{P(X|W, \theta) P(W|\theta) P(\theta)} \frac{q(\theta|\vartheta)}{q(\vartheta|\theta)}. \tag{5}$$

7: **Backward pass:** Simulate $v_{|W|} \sim \mathtt{b}_{|W|}^{\theta'}(\cdot)$, where $\mathtt{b}_{|W|}^{\theta'}(s) \propto \mathtt{f}_{|W|}^{\theta'}(s) \cdot \ell_{|W|+1}(s) \quad \forall s \in \mathcal{S}$.

for $i = (|W|-1) \to 0$ **do:** $v_i \sim \mathtt{b}_i^{\theta'}(\cdot)$, where $\mathtt{b}_i^{\theta'}(s) \propto \mathtt{f}_i^{\theta'}(s) \cdot B_{s v_{i+1}}(\theta') \cdot \ell_{i+1}(s) \; \forall s \in \mathcal{S}$.

8: Set $s_0' = v_0$. Let T' be the set of times in W when V changes state. Define S' as the corresponding set of state values. Return (s_0', S', T', θ').

The resulting MCMC algorithm updates θ with the MJP trajectory integrated out, and by instantiating less 'missing' information, can be expected to mix better. This can

For Peer Review Only

be quantified by the so-called Bayesian fraction of missing information (Liu, 1994; Papaspiliopoulos et al., 2007). We note that even with the state values (v_0, V) marginalized out, θ is updated *conditioned on* W. The distribution of W depends on θ: W follows a rate-$\Omega(\theta)$ Poisson process. This dependence manifests in the $P(W|\theta)$ and $P(W|\vartheta)$ terms in equation (5). The fact that the MH-acceptance involves the probability of the observations X is inevitable, however the $P(W|\theta)$ term is an artifact of the computational algorithm of Rao-Teh. In our experiments, we show that this term significantly hurts acceptance probabilities and mixing. For a given θ, $|W|$ is Poisson distributed with mean and variance $\Omega(\theta)$. If the proposed ϑ is such that $\Omega(\vartheta)$ is half $\Omega(\theta)$, then the ratio $P(W|\vartheta)/P(W|\theta)$ will be small, and ϑ is unlikely to be accepted. The next section describes our main algorithm that gets around this.

5 An improved Metropolis-Hasting algorithm

The algorithm we propose symmetrizes the probability of W under the old and new parameters, so that $P(W|\theta)$ disappears from the acceptance ratio. Now, the probability of accepting a proposal ϑ will depend only on the prior probabilities of θ and ϑ, as well as how well they both explain the data given W. This is in contrast to the previous algorithm, where one must also factor in how well each parameter explains the current value of the grid W. This results in a MCMC sampler that mixes significantly more rapidly. Since we need not account for the probabilities $P(W|\theta)$, we also have a simpler MCMC scheme.

As before, the MCMC iteration begins with (s_0, S, T, θ). Instead of simulating the thinned events U like earlier algorithms, we *first* generate a new parameter ϑ from some distribution $q(\vartheta|\theta)$. Treat this as an auxiliary variable, so that the augmented space now is $(s_0, S, T, \theta, \vartheta)$. Define a function $\Omega(\theta, \vartheta) > \max_s A_s(\theta)$ that is symmetric in its arguments (the number of arguments will distinguish $\Omega(\cdot,\cdot)$ from $\Omega(\cdot)$ of the earlier sections). Two examples are $\Omega(\theta, \vartheta) = \kappa \max_s A_s(\theta) + \kappa \max_s A_s(\vartheta)$, for $\kappa \geq 1$, and $\Omega(\theta, \vartheta) = \kappa \max\left(\max_s A_s(\theta), \max_s A_s(\vartheta)\right)$, for $\kappa > 1$.

We will treat the path (s_0, S, T) as simulated by uniformization, but now with the dominating Poisson rate equal to $\Omega(\theta, \vartheta)$ instead of $\Omega(\theta)$. The transition matrix $B(\theta, \vartheta)$

For Peer Review Only

of the embedded Markov chain is $B(\theta, \vartheta) = I + \frac{1}{\Omega(\theta,\vartheta)} A(\theta)$, so that the resulting trajectory (s_0, S, T) will still be a realization from a MJP with rate-matrix $A(\theta)$.

Following the Rao-Teh algorithm, the conditional distribution of the thinned events U given $(s_0, S, T, \theta, \vartheta)$ is a piecewise-constant Poisson with rate $\Omega(\theta, \vartheta) - A_{S(t)}(\theta), t \in [0, t_{end})$. This reconstructs the set $W = U \cup T$, and as we saw (see also Rao and Teh, 2013), $P(W|\theta, \vartheta)$ is a homogeneous Poisson process with rate $\Omega(\theta, \vartheta)$. Having imputed W, discard the state values, so that the MCMC state space is (W, θ, ϑ). Now, propose swapping θ with ϑ. From the symmetry of $\Omega(\cdot, \cdot)$, the Poisson grid W has the same probability both before and after this proposal, so unlike the previous scheme, the ratio equals 1. This simplifies computation, and as suggested in the previous section, can significantly improve mixing. An acceptance probability of $\min\left(1, \frac{P(X|W,\vartheta,\theta)P(\vartheta)q(\theta|\vartheta)}{P(X|W,\theta,\vartheta)P(\theta)q(\vartheta|\theta)}\right)$ targets the conditional $P(W, \theta, \vartheta|X) \propto P(\theta)q(\vartheta|\theta)P(W, X|\theta, \vartheta)$. The terms $P(X|\vartheta)$ and $P(X|\theta)$ can be calculated from the forward pass of FFBS, and after accepting or rejecting the proposal, a new trajectory is sampled by completing the backward pass. Finally, the thinned events and auxiliary parameter are discarded. Algorithm 4 and figure 2 outline the details of these steps.

For Peer Review Only

Algorithm 4 Symmetrized MH for parameter inference for MJPs

Input: The observations X, the MJP path (s_0, S, T), MJP parameters θ and π_0.

Output: A new MJP trajectory (s_0', S', T'), new MJP parameters θ'.

1: **Sample** $\vartheta \sim q(\cdot|\theta)$, and set $\Omega \doteq \Omega(\theta, \vartheta)$ for some symmetric $\Omega(\theta, \vartheta) > \max_s A_s(\theta)$.

2: **Simulate the thinned times** U from a rate-$(\Omega - A_{S(\cdot)}(\theta))$ Poisson process:

$$U \sim \text{PoissProc}(\Omega - A_{S(t)}(\theta)), \quad t \in [0, t_{end}).$$

3: Set $W = T \cup U$ and discard (s_0, S). Define $\tilde{W} = 0 \cup W \cup t_{end}$.

4: **Forward pass:** Set $B(\theta, \vartheta) = I + \frac{A(\theta)}{\Omega(\theta,\vartheta)}$ and $\mathtt{f}_0^{\theta,\vartheta}(\cdot) = \pi_0$. Recall $\ell_i(\cdot)$ from equation (1).

for $i = 1 \to |W|$ **do:** $\mathtt{f}_i^{\theta,\vartheta}(s') = \sum_{s \in \mathcal{S}} \mathtt{f}_{i-1}^{\theta,\vartheta}(s) \cdot \ell_i(s) \cdot B_{ss'}(\theta, \vartheta), \quad \forall s' \in \mathcal{S}.$

Similarly, for $B(\vartheta, \theta) = I + \frac{A(\vartheta)}{\Omega(\vartheta,\theta)}$, calculate $\mathtt{f}_i^{\vartheta,\theta}(\cdot)$ for all elements of W.

5: **Swap** (θ, ϑ) with probability $1 \wedge \frac{P(X|W,\vartheta,\theta)P(\vartheta)q(\theta|\vartheta)}{P(X|W,\theta,\vartheta)P(\theta)q(\vartheta|\theta)}$. Write the new parameters as (θ', ϑ').

Here $P(X|W, \theta, \vartheta) = \sum_{s \in \mathcal{S}} \mathtt{f}_{|W|}^{\theta,\vartheta}(s)\ell_{|W|+1}(s)$, $P(X|W, \vartheta, \theta) = \sum_{s \in \mathcal{S}} \mathtt{f}_{|W|}^{\vartheta,\theta}(s)\ell_{|W|+1}(s)$.

6: **Backward pass:** Simulate $v_{|W|} \sim \mathtt{b}_{|W|}^{\theta',\vartheta'}(\cdot)$, where $\mathtt{b}_{|W|}^{\theta',\vartheta'}(s) \propto \mathtt{f}_{|W|}^{\theta',\vartheta'}(s) \cdot \ell_{|W|+1}(s) \;\; \forall s \in \mathcal{S}$.

for $i = (|W|-1) \to 0$ **do:** $v_i \sim \mathtt{b}_i^{\theta',\vartheta'}(\cdot)$, where $\mathtt{b}_i^{\theta',\vartheta'}(s) \propto \mathtt{f}_i^{\theta',\vartheta'}(s) \cdot B_{sv_{i+1}}(\theta', \vartheta') \cdot \ell_{i+1}(s) \; \forall s \in \mathcal{S}$.

7: Set $s_0' = v_0$. Let T' be the set of times in W when V changes state. Define S' as the corresponding set of state values. Return (s_0', S', T', θ').

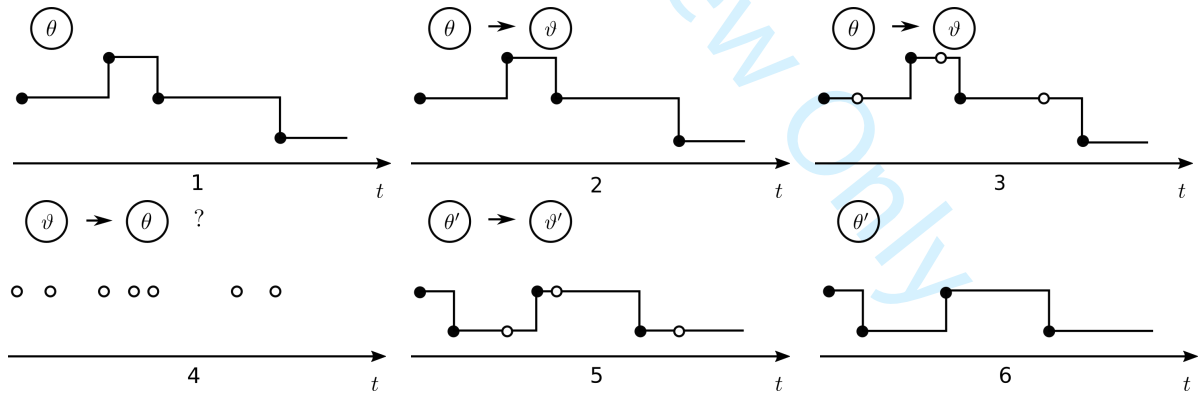

Figure 2: Symmetrized MH algorithm: Steps 1-3: Starting with a trajectory and parameter θ, simulate an auxiliary parameter ϑ, and then the thinned events U from a rate $\Omega(\theta, \vartheta) - A_{S(\cdot)}$ Poisson process. Step 4: Discard state values, and propose swapping θ and ϑ. Step 5: Run a forward pass to accept or reject this proposal, calling the new parameters (θ', ϑ'). Use these to simulate a new trajectory. Step 6: Discard ϑ' and the thinned events.

For Peer Review Only

Proposition 1. *The sampler described in Algorithm 4 has the posterior distribution* $P(\theta, s_0, S, T|X)$ *as its stationary distribution.*

Proof. Consider a realization (θ, s_0, S, T) from the posterior distribution $P(\theta, s_0, S, T|X)$. An iteration of the algorithm first simulates ϑ from $q(\vartheta|\theta)$. By construction, the marginal distribution over all but the last variable in the set $(\theta, s_0, S, T, \vartheta)$ is still the posterior.

The algorithm next simulates U from a Poisson process with rate $\Omega(\theta, \vartheta) - A_{S(\cdot)}(\theta)$. Write $W = T \cup U$. The random grid W consists of the actual and thinned candidate transition times, and is distributed according to a rate-$\Omega(\theta, \vartheta)$ Poisson process (Proposition 2 in Rao and Teh (2013)). Thus, the triplet (W, θ, ϑ) has probability proportional to $P(\theta)q(\vartheta|\theta)\text{PoissProc}(W|\Omega(\theta, \vartheta))P(X|W, \theta, \vartheta)$. Next, the algorithm proposes swapping θ and ϑ with W fixed (a deterministic proposal), and accepts with MH-acceptance probability

$$\texttt{acc} = 1 \wedge \frac{P(\vartheta)q(\theta|\vartheta)P(X|W, \vartheta, \theta)}{P(\theta)q(\vartheta|\theta)P(X|W, \theta, \vartheta)} = 1 \wedge \frac{P(\vartheta)q(\theta|\vartheta)\text{PoissProc}(W|\Omega(\vartheta, \theta))P(X|W, \vartheta, \theta)}{P(\theta)q(\vartheta|\theta)\text{PoissProc}(W|\Omega(\theta, \vartheta))P(X|W, \theta, \vartheta)},$$

where we exploit the symmetry of $\Omega(\cdot, \cdot)$. Write the new parameters as (θ', ϑ').

This MH step has stationary distribution over (W, θ', ϑ') proportional to $P(\theta')q(\vartheta'|\theta)$ $\text{PoissProc}(W|\Omega(\theta', \vartheta'))P(X|W, \theta', \vartheta')$, so that the triplet (W, θ', ϑ') has the same distribution as (W, θ, ϑ). The algorithm uses $B(\theta', \vartheta')$ to make a backward pass through W, simulating state values on W from the conditional of a Markov chain with transition matrix $B(\theta', \vartheta')$ given observations X. Dropping the self-transition times results in $(\theta', s_0', S', T', \vartheta')$. From uniformization (see also Lemma 1 in Rao and Teh (2013)), the trajectory (s_0', S', T') is distributed according to the conditional of a rate-$A(\theta')$ MJP given observations X. Finally, dropping ϑ' results in (θ', s_0', S', T') from the posterior given X, proving stationarity. □

6 Related work

Our paper modifies the algorithm from Rao and Teh (2013) to include parameter inference. That algorithm requires a uniformization rate $\Omega(\theta) > \max_s A_s(\theta)$, and empirical results from that paper suggest $\Omega(\theta) = 2\max_s A_s(\theta)$. The uniformization rate $\Omega(\theta, \vartheta)$ in our algorithm includes a proposed parameter ϑ, must be symmetric in both arguments and must be greater than both $\max_s A_s(\theta)$ and $\max_s A_s(\vartheta)$. A natural and simple setting is

For Peer Review Only

$\Omega(\theta, \vartheta) = \max_s A_s(\theta) + \max_s A_s(\vartheta)$. When θ is known, our algorithm has ϑ equal to θ (i.e. the 'proposed' ϑ equals θ), and our uniformization rate reduces to $2 \max_s A_s(\theta)$. This provides a principled motivation for the particular choice of Ω in Rao and Teh (2013).

Of course, we can consider other choices, such as $\Omega(\theta, \vartheta) = \kappa(\max A_i(\theta) + \max A_i(\vartheta))$ for $\kappa > 1$. These result in more thinned events, and so more computation, with faster MCMC mixing. We study the effect of κ in our experiments, but find the smallest setting of $\kappa = 1$ performs best. It is also possible to have non-additive settings for $\Omega(\theta, \vartheta)$, for example, $\Omega(\theta, \vartheta) = \kappa \max(\max_i A_i(\theta), \max A_i(\vartheta))$ for some $\kappa > 1$. We investigate this too.

A key idea in our paper, as well as Rao and Teh (2013), is to impute the random grid of candidate transition times W every MCMC iteration. Conditioned on W, the MJP trajectory follows an HMM with transition matrix B. By running the FFBS algorithm over W, we can marginalize out the states associated with W, and calculate the marginal $P(X|W, \theta)$. Another approach to parameter inference that integrates out state values follows Fearnhead and Sherlock (2006). This algorithm makes a sequential forward pass through all *observations* X (rather than W). Unlike with W fixed, one cannot a priori bound the number of transitions between two successive observations, so that Fearnhead and Sherlock (2006) have to use matrix exponentials of A (rather than just B) to calculate transition probabilities. The resulting algorithm is cubic, rather than quadratic in the number of states, and the number of expensive matrix exponentiations needed scales with the number of observations, rather than the number of transitions. Further, matrix exponentiation results in a dense matrix, so that Fearnhead and Sherlock (2006) cannot exploit sparsity in the transition matrix. In our framework, $B = I + \frac{1}{\Omega} A$ inherits sparsity present in A. Thus if A is tri-diagonal, our algorithm is *linear* in the number of states.

A second approach to marginalizing out state information is particle MCMC (Andrieu et al., 2010). This algorithm, described in section 11.3 in the supplementary material, uses particle filtering to get an unbiased estimate of $P(X|\theta)$. Plugging this estimate into the MH acceptance probability results in an MCMC sampler that targets the correct posterior, however the resulting scheme does not exploit the Markovian structure of the MJP the way FFBS can. In particular, observations that are informative of the MJP state can result

For Peer Review Only

in marginal probability estimates that have large variance, resulting in slow mixing. By contrast, given W, FFBS can compute the marginal probability $P(X|W,\theta)$ *exactly*.

The basic idea of marginalizing out information to accelerate MCMC convergence is formalized by the idea of the Bayesian fraction of missing information (Liu, 1994). In this context, papers such as Papaspiliopoulos et al. (2007); Yu and Meng (2011) have studied MCMC algorithms for hierarchical latent variable models. The Gibbs sampler of algorithm 2 can be viewed as operating on a centered parametrization (Papaspiliopoulos et al., 2007) or sufficient augmentation (Yu and Meng, 2011) of a hierarchical model involving the parameter θ, the latent variables (v_0, V, W) and the observations X. These papers then suggest noncentered parametrizations or ancillary augmentations, which in our context correspond to simulating θ, W, and an *independent* set of $(|W|+1)$ i.i.d. uniform random variables Q. Through a sequence of inverse-cdf transforms, the state values (v_0, V) are then written as a deterministic function of Q and θ: $(v_0, V) = f_\theta(Q)$, after which the observations X are produced. Now, proposing a new parameter ϑ automatically proposes a new set of state variables $(v_0', V') = f_\vartheta(Q)$, so that problem of path-parameter coupling is avoided. A similar idea could also be used to avoid couplng between θ and the Poisson process W. However now, updating Q given θ and (v_0, V, W) raises significant challenges to mixing. By contrast, our approach marginalizes out the variables (v_0, V) (or Q), and will mix significantly faster. Nevertheless, results from the literature on NCPs can suggest further improvements to our approach, and give guidance about conditions under which approaches like ours outperform centered parametrisations like algorithm 2, or when a mixture of centered and non-centered updates could be useful (Yu and Meng, 2011).

Our approach of first simulating ϑ, and then simulating W from a Poisson process whose rate is symmetric in θ and ϑ is related to Neal (2004). In that work, to simulate from an 'energy' model $P(x,y) \propto \exp(-E(x,y))$, the author proposes a new parameter x^*, and then updates y via intermediate transitions to be symmetric in x and x^*, before proposing to swap x and x^*. Our approach exploits the specific structure of the Poisson and Markov jump processes to do this directly, avoiding the need for any tempered transitions.

Our algorithm is also related to work on MCMC for doubly-intractable distributions.

For Peer Review Only

Algorithms like Møller et al. (2006); Murray et al. (2006); Andrieu and Roberts (2009) all attempt to evaluate an intractable likelihood under a proposed parameter ϑ by introducing auxiliary variables, typically sampled independently under the proposed parameters. For MJPs, this would involve proposing ϑ, generating a new grid W^*, and then using $P(X|W,\theta)$ and $P(X|W^*,\vartheta)$ in the MH acceptance step. This is more involved (with two sets of grids), and introduces additional variance that reduces acceptance rates. While Murray et al. (2006) suggest annealing schemes to try to address this issue, we exploit the uniformization structure to provide a cleaner solution: generate a single set of auxiliary variables that depends symmetrically on both the new and old parameters.

7 Experiments

In the following, we evaluate Python implementations of a number of algorithms, focusing on our contribution, the symmetrized MH algorithm (algorithm 4), and as well as the naïve MH algorithm (algorithm 3). We evaluate different variants of these algorithms, corresponding to different uniformizing Poisson rates. For naïve MH, we set $\Omega(\theta) = \kappa \max_s A_s(\theta)$ with κ equal to $1.5, 2$ and 3 (here κ must be greater than 1), while for symmetrized MH, where the uniformizing rate depends on both the current and proposed parameters, we consider $\Omega(\theta,\vartheta) = \kappa(\max A(\theta) + \max A(\vartheta))$ ($\kappa = 1$ and 1.5), and $\Omega(\theta,\vartheta) = 1.5\max(\max A(\theta), \max A(\vartheta))$. We evaluate two other baselines: Gibbs sampling (algorithm 2), and particle MCMC (Andrieu et al., 2010, see also section 11.3 in the appendix). Gibbs sampling involves a uniformization step to update the MJP trajectory (step 1 in algorithm 2), for which we use $\Omega(\theta) = \kappa \max_s A_s(\theta)$ for $\kappa = 1.5, 2, 3$. Unless specified, our results were obtained from 100 independent MCMC runs, each of 10000 iterations. We found particle MCMC to be more computationally intensive, and limited each run to 3000 iterations, the number of particles being $5, 10$ and 20.

For each run of each MCMC algorithm, we calculated the effective sample size (ESS) of the posterior samples of the MJP parameters using the R package `rcoda` (Plummer et al., 2006). This estimates the number of independent samples returned by the MCMC algorithm, and dividing this by the runtime of a simulation gives the ESS per unit time (ESS/sec). We used this to compare different samplers and different parameter settings.

For Peer Review Only

7.1 A simple synthetic MJP

Consider an MJP with a uniform distribution over states at time 0, and with transitions between states i and j having rate $\alpha \exp(-\beta/(i+j))$, for two parameters $(\alpha, \beta) \stackrel{\text{def}}{=} \theta$. We consider three settings: 3 states (figure 3), 5 states, and 10 states. We place Gamma(α_0, α_1), and Gamma(β_0, β_1) priors on the parameters α and β, with $(\alpha_0, \alpha_1, \beta_0, \beta_1)$ having values $(3, 2, 5, 2)$ respectively. For each run, we draw random parameters from the prior to construct a transition matrix A, and simulate an MJP trajectory. We simulate observations uniformly at integer values on the time interval $[0, 20]$. Each observation is Gaussian distributed with mean equal to the state at that time, and variance equal to 1. For the MH proposal, we used a lognormal distribution centered at the current parameter value, with variance σ^2 whose effect we study.

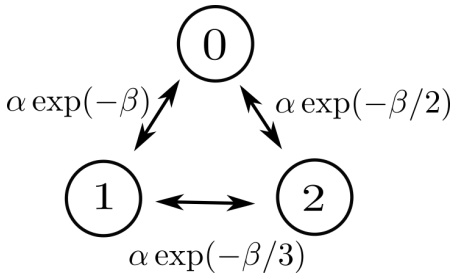

Figure 3: A 3-state MJP with exponentially decaying rates

Results: Figure 4 shows the MCMC estimates of the posterior distribution over α, $P(\alpha|X)$ from the Gibbs sampler as well as our symmetrized MH sampler. Visually these agree, and we quantify this by running a Kolmogorov-Smirnov two-sample test using 1000 samples from each algorithm: this returns a p-value of 0.5085, clearly failing to reject the null hypothesis that both samples come from the same distribution. The same is true for β, though we do not include it here. The figure also shows the average acceptance probabilities for the two MH samplers: we see that for the same proposal distribution, symmetrization significantly improves acceptance probability. This shows the benefit of eliminating the $P(W|\theta)$ terms from the acceptance probability (we will investigate this further). Figure 5 shows traceplots and autocorrelation plots for α from the symmetrized MH and Gibbs samplers. Clearly, our sampler mixes much more efficiently than Gibbs, with naïve MH (included in the supplementary material) worse than both.

For Peer Review Only

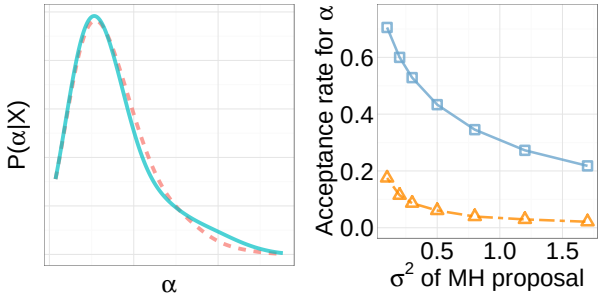

Figure 4: (Left) posterior $P(\alpha|X)$ from Gibbs (dashed line) and symmetrized MH (solid line) for the synthetic model. (Right) acceptance probabilities of α for symmetrized (squares) and naïve (triangles) MH.

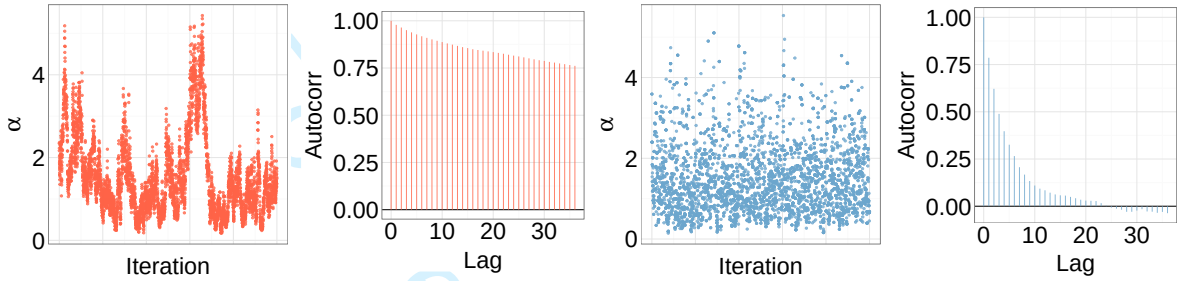

Figure 5: Trace and autocorrelation plots for Gibbs (left two panels) and symmetrized MH (right two panels). All plots are for the synthetic model with 10 states.

To quantify performance, figure 6 plots the ESS/sec in the top row, and ESS per 1000 samples in the bottom row for α and β. The left two columns consider α and β for MJPs with 3 states, and the right two, with 10 states. We include results for 5 states in the supplementary material, the conclusions are the same. For each plot, we vary the scale-parameter σ^2 of the log-normal proposal $q(\vartheta|\theta)$, and look at its effects on ESS/s and ESS. Note that the conditional over parameters given trajectory is not conjugate, so that the Gibbs sampler is really a Metropolis-within-Gibbs (MWG) sampler with an associated lognormal proposal distribution parameterized by σ^2.

We see that our symmetrized MH algorithm, shown with blue squares, is significantly more efficient than the baselines over a wide range of σ^2 values, including the natural choice of 1. Among the baselines, Gibbs (red circles) does better than naïve MH (yellow triangles), confirming that the dependency of the Poisson grid on the MJP parameters (as indicated in figure 4) does indeed slow down mixing. This, coupled with the fact that MWG tends to have higher MH acceptance than naïve MH results in Gibbs having superior performance. Our symmetrized MH avoids this problem at no additional computational cost. Particle MCMC (black diamonds) has the worst performance.

For Peer Review Only

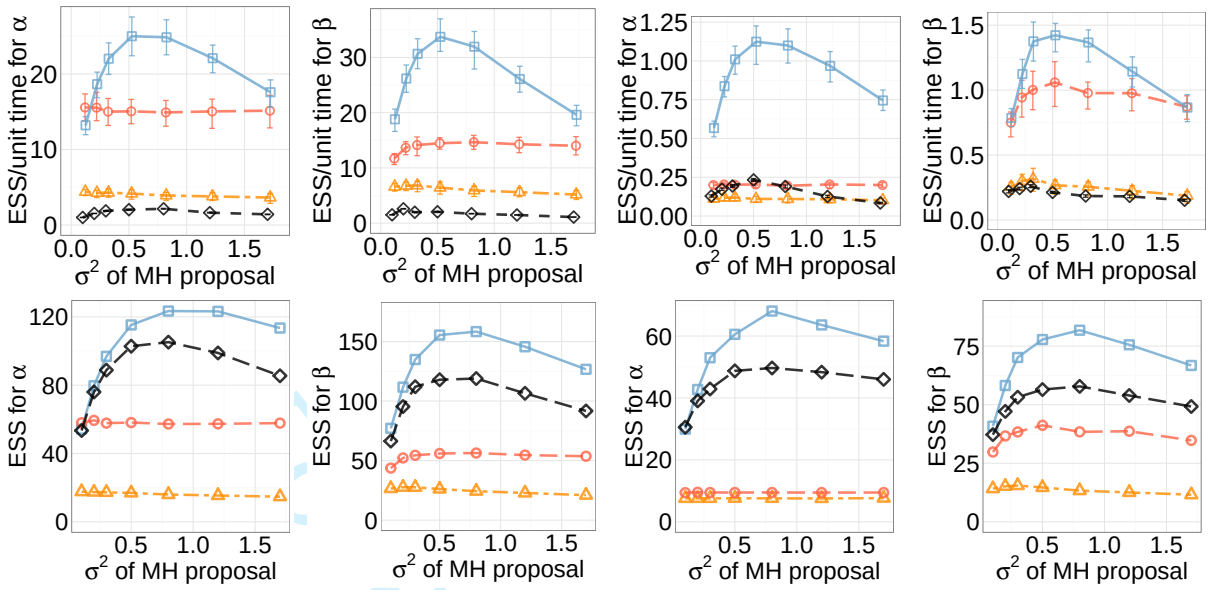

Figure 6: ESS/sec (top row) and ESS per 1000 samples (bottom row) of different algorithms on the synthetic model. The left two panels are α and β for 3 states, the right two, for 10 states. Blue squares, yellow triangles, red circles and black diamonds are the symmetrized MH, naïve MH, Gibbs and particle MCMC algorithm.

Among the three setting of our algorithm, the simple additive setting (squares) does best, slightly better than the max-of-max setting (circles). The additive setting with a multiplicative factor of 1.5 (triangles) does worse than both the additive choice with $\kappa = 1$ and the max-of-max choice but still better than the other algorithms. The results in figure 6 for 10 states shows that ESS is slightly lower, and thus mixing is slightly poorer for all samplers. This, coupled with greater computational cost per iteration results in a drop in ESS/s across all algorithms, compared with 3 states. Our symmetrized MH algorithm still outperforms the other samplers, and we observe the same pattern of relative performance for different settings of our sampler (figure 7), with a uniformization rate of $\Omega(\theta, \vartheta) = \max_s A_s(\theta) + \max_s A(\vartheta)$ giving the best performance.

For Peer Review Only

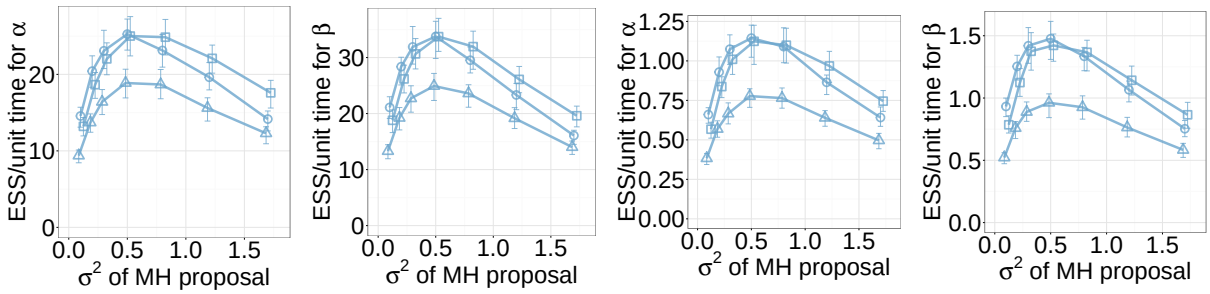

Figure 7: ESS/sec of symmetrized MH for different choices of $\Omega(\theta, \vartheta)$ for the synthetic model. The left two panels are α and β for 3 states, and the right two for 10 states. Squares, circles and trianges correspond to $\Omega(\theta, \vartheta)$ set to $(\max_s A_s(\theta) + \max_s A_s(\vartheta))$, $\max(\max_s A_s(\theta), \max_s A_s(\vartheta))$ and $1.5(\max_s A_s(\theta) + \max_s A_s(\vartheta))$.

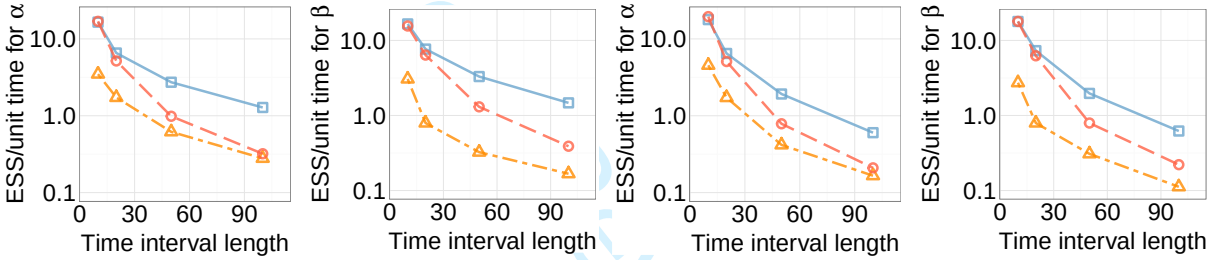

Figure 8: Time interval vs ESS/sec for the synthetic MJP. The left two plots are for α and β, with the number of observations fixed; in the right two, this grows linearly with the interval length. Blue squares, yellow triangles and red circles curves are the symmetrized MH, naïve MH and Gibbs algorithm.

In figure 8, we plot ESS per unit time as the observation interval t_{end} increases. We consider the 3-state MJP, and as before there are 19 observations uniformly located over a time interval $(0, t_{end})$. We consider four settings, with t_{end} equal to $10, 20, 50, 100$. For each, we compare our symmetrized MH sampler (with κ set to 1) with the naïve MH and Gibbs samplers (with κ set to 2). While the performance of the Gibbs sampler is comparable with our symmetrized algorithm for the smallest value of t_{end}, its performance is considerably worse for longer time-intervals. This is the limitation of Gibbs sampling that motivated this work: when updating θ conditioned on the MJP trajectory, longer time intervals result in stronger coupling between MJP path and parameters (figure 1), and thus poorer mixing.

For Peer Review Only

The performance of the naïve sampler demonstrates that it is not sufficient just to integrate out the state values of the trajectory, we also have to get around the coupling between the Poisson grid and the parameters. Our symmetrized MH-algorithm allows this.

To the right of figure 8, we plot results from a similar experiment. Now, instead of keeping the number of measurements fixed as we increase the observation interval, we keep the observation *rate* fixed at one observation every unit interval of time, so that longer observation intervals have larger number of observations. The results are similar to the previous case: Gibbs sampling performs well for small observation intervals, with performance degrading sharply for larger intervals.

7.2 The Jukes and Cantor (JC69) model

The Jukes and Cantor (JC69) model (Jukes and Cantor, 1969) is a popular model of DNA nucleotide substitution. We write its state space as $\{0, 1, 2, 3\}$, representing the four nucleotides $\{A, T, C, G\}$. The model has a single parameter α, representing the rate at which the system transitions between any pair of states. Thus, the rate matrix A is given by $A_i = -A_{i,i} = 3\alpha, A_{i,j} = \alpha, i \neq j$. We place a Gamma$(3, 2)$ prior on the parameter α. Figures 9(b) and (c) compare different samplers: we again see that the symmetrized MH samplers comprehensively outperforms all others. Part of the reason why the difference is so dramatic here is because now a *single* paramter $\alpha \stackrel{\text{def}}{=} \theta$ defines the transition matrix, implying a stronger coupling between MJP path and parameter. We point out that for Gibbs sampling, the conditional distribution over θ is conjugate to the Gamma prior. We can thus simulate directly from this distribution without any MH proposal (hence its performance remains fixed along the x-axis). Despite this, its performance is worse than our symmetrized algorithm. Particle MCMC performs worse than all the algorithms, and we do not include it in our plots. Figure 9(d) compares different settings of $\Omega(\theta, \vartheta)$ for our sampler: again, the simple additive setting $\Omega(\theta, \vartheta) = \max_s A_s(\theta) + \max_s A_s(\vartheta)$ does best.

For Peer Review Only

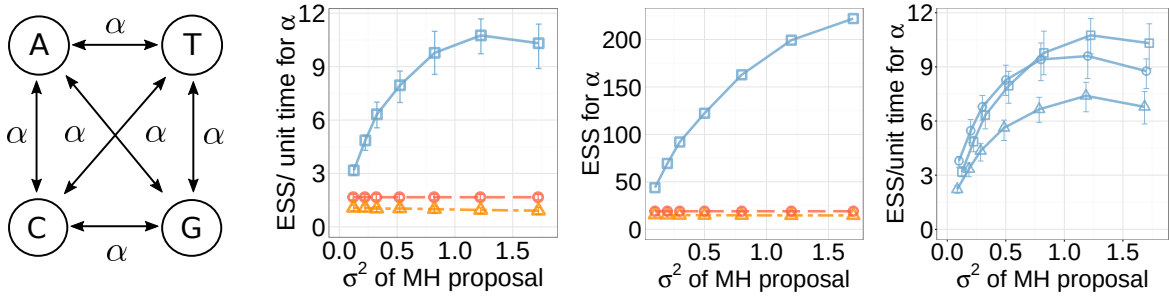

Figure 9: The leftmost panel is the Jukes-Cantor (JC69) model. The next two panels from left to right are ESS/sec and ESS per 1000 samples for this. Blue squares, yellow triangles and red circles are the symmetrized MH, naïve MH and Gibbs algorithm. The rightmost panel looks at different settings of the symmetrized MH algorithm, with squares, circles and triangles corresponding to $\Omega(\theta, \vartheta)$ set to $(\max_s A_s(\theta) + \max_s A_s(\vartheta))$, $\max(\max_s A_s(\theta), \max_s A_s(\vartheta))$ and $1.5(\max_s A_s(\theta) + \max_s A_s(\vartheta))$.

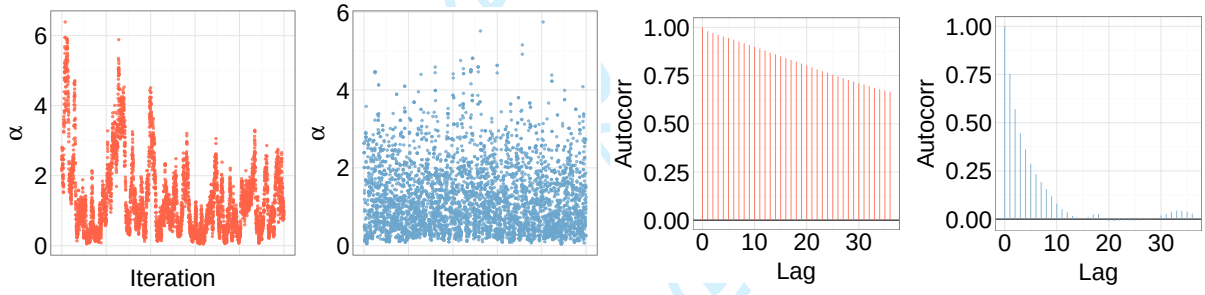

Figure 10: Trace and autocorrelation plots of α for the JC69 model. Left two panels are for Gibbs and the right two for the symmetrized MH algorithm.

Figure 10 plots MCMC diagnostics for the Gibbs and symmetrized MH sampler, confirming the previous findings. Both agree on the posterior $P(\alpha|X)$ (figure 11(a)), with a two-sample Kolmogorov-Smirnov test giving a p-value of 0.97. Figure 11(b) plots the average MH acceptance probabilities for the naïve and symmetrized MH samplers for different settings of the proposal distribution, again we see that the former has lower acceptance rates because of the $P(W|\theta)$ terms. Figures 11 (c) and (d) plot the ESS per unit time for the different samplers as t_{end} increases. The left plot keeps the number of observations fixed, while the right keeps the observation rate fixed. Once again we see that our proposed algorithm 1) performs best over all interval lengths, and 2) suffers a performance degradation with interval length that is much milder than the other algorithms.

For Peer Review Only

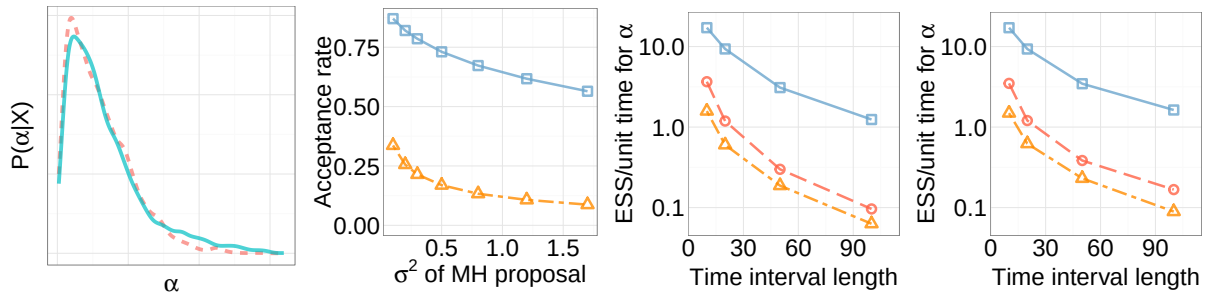

Figure 11: (a) Posterior $P(\alpha|X)$ in the JC69 model for Gibbs (dashed) and symmetrized MH (continuous). (b) MH acceptance rates for naïve and symmetrized MH. (c) and (d): ESS/sec against t_{end} for $\kappa = 2$ with: (c) number of observations fixed, and (d) observation rate fixed. Squares, triangles and circles are symmetrized MH, naïve MH and Gibbs.

7.3 An immigration model with finite capacity

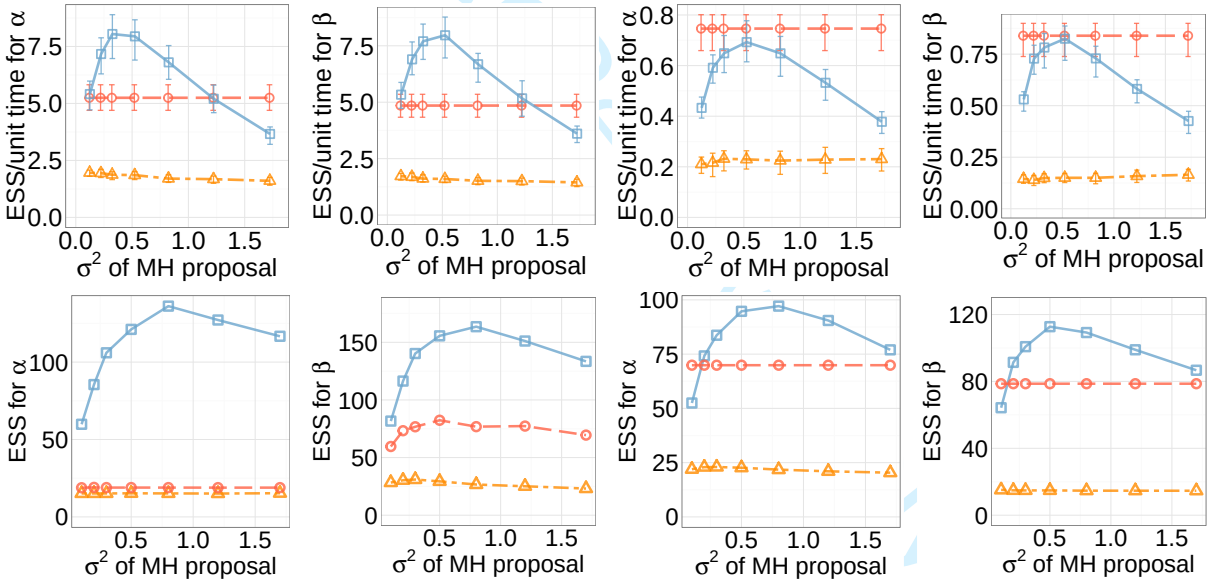

Figure 12: ESS/sec (top row) and ESS per 1000 samples (bottom row) for the immigration model. The left two columns are α and β for 3 states, and the right two, for 10 states. Squares, triangles and circles are symmetrized MH, naïve MH, and Gibbs algorithm.

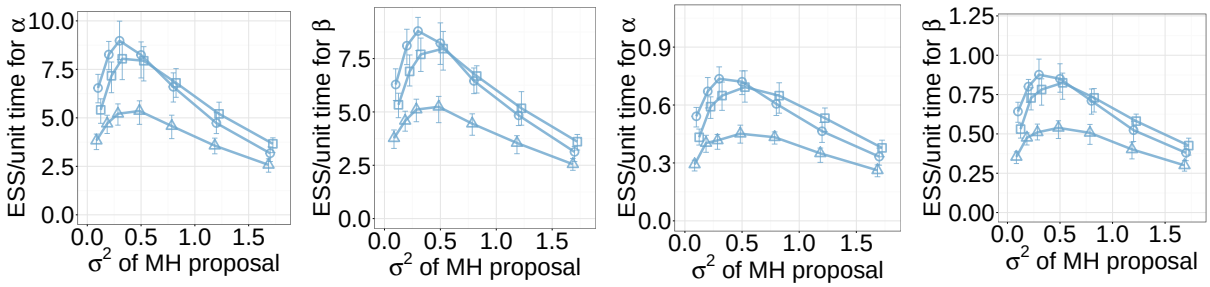

Figure 13: ESS/sec for symmetrized MH for the immigration model for different settings of $\Omega(\theta, \vartheta)$. The left two columns are for α and β with 3 states, and the right two, with 10. Squares, circles and triangles correspond to $\Omega(\theta, \vartheta)$ set to $(\max_s A_s(\theta) + \max_s A_s(\vartheta))$, $\max(\max_s A_s(\theta), \max_s A_s(\vartheta))$ and $1.5(\max_s A_s(\theta) + \max_s A_s(\vartheta))$.

Next, we consider an M/M/N/N queue (Gross et al., 2011). The state space of this stochastic process is $\{0, 1, 2, 3, \cdots, N-1\}$ giving the number of customers/jobs/individuals in a system/population. Arrivals follow a rate-α Poisson process, moving the process from state i to $i+1$ for $i < N$. The system has a capacity of N, so any arrivals when the current state is N are discarded. Service times or deaths are exponentially distributed, with a rate that is now state-dependent: the system moves from i to $i-1$ with rate $i\beta$.

We follow the same setup as the first experiment: for $(\alpha_0, \alpha_1, \beta_0, \beta_1)$ equal to $(3, 2, 5, 2)$, we place Gamma(α_0, α_1), and Gamma(β_0, β_1) priors on α, β. These prior distributions are used to sample transition matrices A, which, along with a uniform distribution over initial states, are used to generate MJP trajectories. We observe these at integer-valued times according to a Gaussian likelihood. We again consider three settings: $3, 5$ and 10 states, with results from 5 steps included in the supplementary material.

Figure 12 plots the ESS per unit time (top row) as well as ESS per 1000 samples (bottom row) for the parameters α and β, again as we change the variance of the proposal kernel. The left two columns show these for α and β for the MJP state-space having size 3, while the right two columns show these for size 10. Our symmetrized MH algorithm does best for dimensions 3 and 5 (shown in the supplement), although now Gibbs sampling performs best for dimensionality 10 (although there is no significant different between the best proposal variance for our sampler and the Gibbs sampler). The Gibbs sampler

For Peer Review Only

performs so well partly because the conditionals over α and β are conjugate, following simple Gamma distributions. Also, unlike the earlier problem, the rate matrix is tri-diagonal, and governed by two parameters, so that path-parameter coupling is now milder.

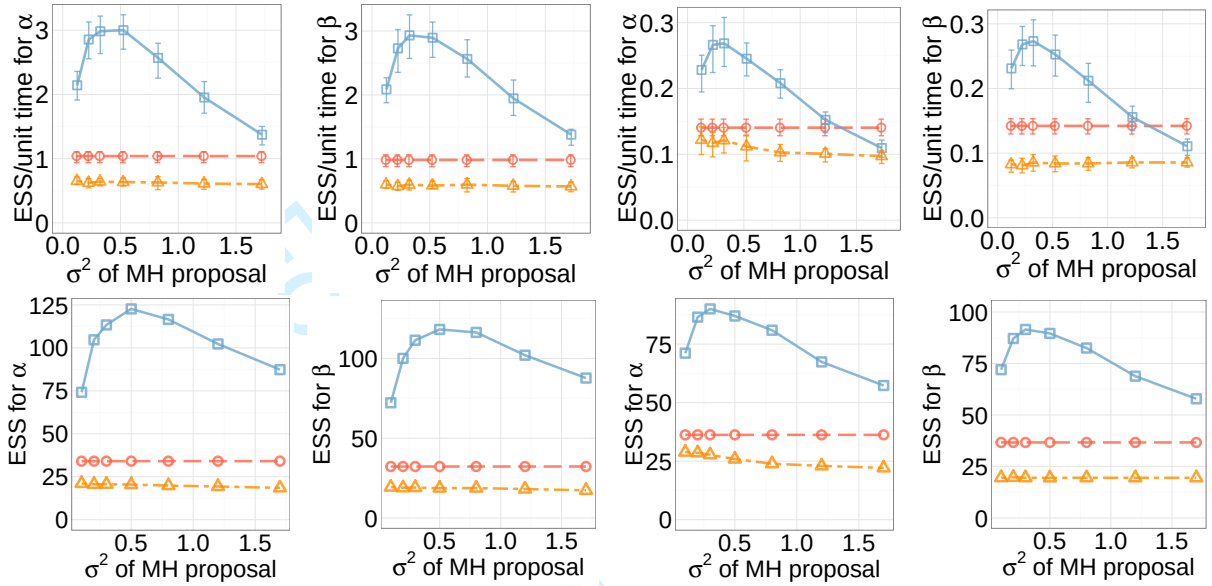

Figure 14: ESS/sec (top row) and ESS per 1000 samples (bottom row) for the time-inhomogeneous immigration model. The left columns are α and β for 3 states, and the right two for 10. Blue squares, yellow triangles and red circles are the symmetrized MH, naïve MH, and Gibbs algorithm.

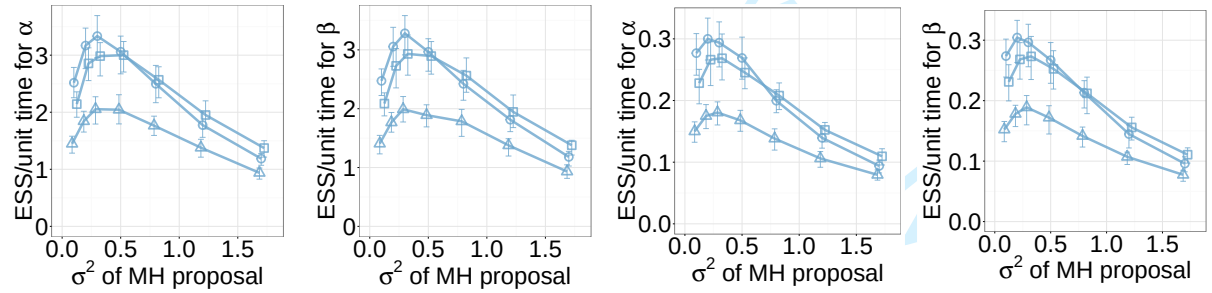

Figure 15: ESS/sec for symmetrized MH for the time-inhomogeneous immigration model for different settings of $\Omega(\theta, \vartheta)$. The left two columns are α and β for 3 states, and the right two for 10. Squares, circles and trianges correspond to $\Omega(\theta, \vartheta)$ set to $(\max_s A_s(\theta) + \max_s A_s(\vartheta))$, $\max(\max_s A_s(\theta), \max_s A_s(\vartheta))$ and $1.5(\max_s A_s(\theta) + \max_s A_s(\vartheta))$.

A time-inhomogeneous immigration model: We extend the previous model to incorporate a known time-inhomogeneity. The arrival and death rates are now no longer

constant, and are instead given by $A_{i,i+1}(t) = \alpha w(t)$ $(i = 0, 1, \cdots, N-1)$ respectively. While it is not difficult to work with sophisticated choices of $w(t)$, we limit ourselves to a simple piecewise-constant $w(t) = \lfloor \frac{t}{5} \rfloor$. Even such a simple change in the original model can dramatically affect the performance of the Gibbs sampler.

Figure 14 plots the ESS per unit time (top row) and ESS per 1000 samples (bottom row) for the parameters α and β. The left two columns show these for this model with capacity 3, and the right two show these for capacity 10. Now, the symmetrized MH algorithm is significantly more efficient, comfortably outperforming all samplers (including the Gibbs sampler) over a wide range of settings. We note that increasing the dimensionality of the state space results in a more concentrated posterior, shifting the optimal setting of the proposal variance to smaller values. Figure 1 shows prior and conditional distributions over α for t_{end} set to 10 and 100, with 3 states.

7.4 Chi site data for Escherichia coli

Finally, we consider a dataset recording positions of a particular DNA motif on the *E. coli* genome. These motifs consist of eight base pairs GCTGGTGG, and are called Chi sites (Fearnhead and Sherlock, 2006). The rates of occurence of Chi sites provide information about genome segmentation, allowing the identification of regions with high mutation or recombination rates. Following Fearnhead and Sherlock (2006), we use this dataset to infer a two-state piecewise-constant segmentation of the DNA strand. We focus on Chi sites along the inner (lagging) strand of the *E. coli* genome. We place an MJP prior over this segmentation, and indexing position along the strand with t, we write this as $\{S(t), t \in [0, 2319.838]\}$. To each state $s \in \{1, 2\}$, we assign a rate λ_s, which together with $S(\cdot)$, defines a piecewise-constant rate function $\lambda_{S(\cdot)}$. We model the Chi-site positions as drawn from a Poisson process with rate $\{\lambda_{S(t)}, t \in [0, 2319.838]\}$, resulting in a Markov-modulated Poisson process (Scott and Smyth, 2003) (see also section 3). MJP transitions from state 1 to state 2 have rate α while transitions from state 2 to state 1 have rate β. We place Gamma$(2, 2)$, Gamma$(2, 3)$, Gamma$(3, 2)$, Gamma$(1, 2)$ priors on α, β, λ_1, λ_2.

We use this setup to evaluate our symmetrized MH sampler along with Gibbs sampling (other algorithms perform much worse, and we do not include them). For our MH proposal

For Peer Review Only

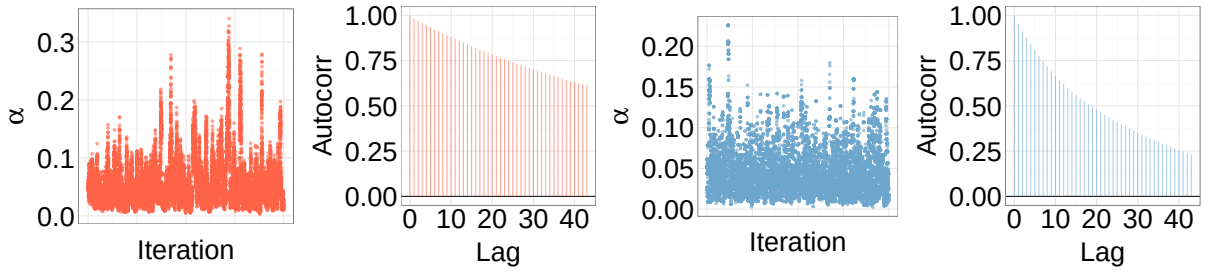

Figure 16: Trace and autocorrelation plots of posterior samples for α for the E. Coli data. The left two plots are the Gibbs sampler and the right two are the symmetrized MH.

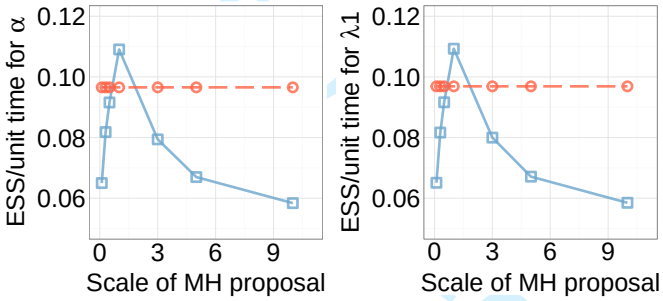

Figure 17: ESS/sec for (α, λ_1) for the E. Coli data. Blue squares are symmetrized MH as we vary the variance of the proposal distribution. The red circles are Gibbs.

distribution, we follow Gelman et al. (2013), and first run 2000 iterations of Gibbs sampling to estimate the posterior covariance of the vector $\theta = (\alpha, \beta, \lambda_1, \lambda_2)$. Call this Σ_θ. Our MH proposal distribution is then $q(\nu|\theta) = N(\nu|\theta, \sigma^2\Sigma_\theta)$ for different settings of σ^2 (we recommend $\sigma^2 = 1$), where we set $\Omega(\theta, \vartheta) = \max_s A_s(\theta) + \max_s A_s(\vartheta)$.

Figure 16 shows trace and autocorrelation plots for the parameter α produced by the Gibbs sampler (left) and our proposed sampler with κ set to 1. We see that this is a fairly hard MCMC sampling problem, however our sampler clearly outperforms Gibbs, which mixes very poorly. Both posterior distributions agreed with each other though, with a two sample-Kolmogorov Smirnov test returning a p-value of 0.1641.

Figure 17 shows the ESS/s for different κ, for parameters (α, λ_1). Both have similar results, and as suggested by the earlier figure, we see that for the typical setting of $\kappa = 1$, our sampler ourperforms Gibbs. In this problem though, Gibbs sampling does outperform our method for large or small κ. This is because a) large or small κ mean the proposal variance is too large or too small, and b) the Gibbs conditionals over the parameters are conjugate for this model. We expect the improvements our method offers to be more robust to the proposal distribution for more complex models without such conditional conjugacy.

For Review Only

8 Geometric ergodicity

We derive conditions under which our symmetrized MH algorithm inherits mixing properties of an 'ideal' sampler that can compute the marginal likelihood $P(X|\theta)$, with the MJP path integrated out. This algorithm proposes a new parameter ϑ from $q(\vartheta|\theta)$, and accepts with probability $\alpha_I(\theta, \vartheta; X) = 1 \wedge \frac{P(X,\vartheta)q(\theta|\vartheta)}{P(X,\theta)q(\vartheta|\theta)}$. The resulting Markov chain has transition probability $P_I(\theta'|\theta) = q(\theta'|\theta)\alpha_I(\theta, \theta'; X) + \left[1 - \int \mathrm{d}\vartheta q(\vartheta|\theta)\alpha_I(\theta, \vartheta; X)\right] \delta_\theta(\theta')$, the first term corresponding to acceptance, and the second, rejection (Meyn and Tweedie, 2009).

Our main result is Theorem 3, which shows that if the ideal MCMC sampler is geometrically ergodic, then so is our sampler in Algorithm 4. Informally, an MCMC algorithm is geometrically ergodic when the total variation distance between the distribution over states and the stationary distribution decreases geometrically with the number of iterations. Meyn and Tweedie (2009) provide more details, as well as sufficient conditions that we exploit in Theorem 3. Geometric ergodicity is an important property of an MCMC chain, guaranteeing that the central limit theorem (CLT) holds for ergodic averages calculated with MCMC samples. Before diving into the proofs, we first state our assumptions,

Assumption 1. *The uniformization rate is set as* $\Omega(\theta, \vartheta) = \Omega(\theta) + \Omega(\vartheta)$*, where* $\Omega(\theta) = k_1 \max_s A_s(\theta) + k_0$*, for some* $k_1 > 1, k_0 > 0$.

Although it is possible to specify broader conditions under which our result holds, for clarity we focus on this case. In our experiments, one of our settings had $k_1 = 1$. We believe our result holds for this case too, but do not prove it. We can drop k_0 if $\inf_\theta \max_s A_s(\theta) > 0$.

Assumption 2. *There exists a positive constant* θ_0 *such that for any* θ_x, θ_y *satisfying* $\|\theta_x\| \geq \|\theta_y\| > \theta_0$*, we have* $\Omega(\theta_x) \geq \Omega(\theta_y)$.

We make this assumption to avoid book-keeping, so $\Omega(\theta)$ increases monotonically with θ.

Definition 1. *Let* π_θ *be the stationary distribution of the MJP with rate-matrix* $A(\theta)$*, and define* $D_\theta = diag(\pi_\theta)$*. Define* $\tilde{A}(\theta) = D_\theta^{-1} A(\theta) D_\theta$*, and the* reversibilization *of* $A(\theta)$ *as* $R_A(\theta) = (A(\theta) + \tilde{A}(\theta))/2$.

For Peer Review Only

This definition is from Fill (1991), who shows that $R_A(\theta)$ is reversible with real eigenvalues, the smallest being 0. The larger its second smallest eigenvalue, the faster the MJP converges to its stationary distribution π_θ. Note that if $A(\theta)$ is reversible, then $R_A(\theta) = A(\theta)$.

Assumption 3. *Write $\lambda_2^{R_A}(\theta)$ for the second smallest eigenvalue of $R_A(\theta)$. There exist $\mu > 0, \theta_1 > 0$ such that for all θ satisfying $\|\theta\| > \theta_1$, we have $\lambda_2^{R_A}(\theta) \geq \mu \max_s A_s(\theta)$ (or equivalently from Assumption 1, $\lambda_2^{R_A}(\theta) \geq \mu\Omega(\theta)$), and $\min_s \pi_\theta(s) > 0$.*

This assumption is the strongest we need, requiring that $\lambda_2^{R_A}(\theta)$ (which sets the MJP mixing rate) grows at least as fast as $\max_s A_s(\theta)$. This is satisfied when θ is bounded, or when, as in our experiments, all elements of $A(\theta)$ grow with θ at similar rates, controlling the relative stability of the least and most stable states. While not trivial, this is reasonable: the MCMC chain over MJP paths will mix well if we can control the mixing of the MJP itself. A less restrictive assumption would also account for the tail behavior of the prior over θ, though we do not do this. To better understand this assumption, recall $B(\theta, \theta') = I + \frac{A(\theta)}{\Omega(\theta,\theta')}$ is the transition matrix of the embedded Markov chain from uniformization, which has the same stationary distribution π_θ as $A(\theta)$. Define the reversibilization $R_B(\theta, \theta')$ of $B(\theta, \theta')$ just as we did $R_A(\theta)$ from $A(\theta)$.

Lemma 2. *Consider $\|\theta\| > \max(\theta_0, \theta_1)$ and θ' such that $\frac{1}{K_0} \leq \frac{\Omega(\theta')}{\Omega(\theta)} \leq K_0$, where K_0 satisfies $(1 + \frac{1}{K_0})k_1 \geq 2$. For all such (θ, θ'), the Markov chain with transition matrix $B(\theta, \theta')$ converges geometrically to stationarity at a rate uniformly bounded away from 0.*

Proof. A little algebra gives $R_B(\theta, \theta') = I + R_A(\theta)/\Omega(\theta, \theta')$. It follows that both R_A and R_B share the same eigenvectors, with eigenvalues satisfying $\lambda_{R_B}(\theta, \theta') = 1 - \frac{\lambda_{R_A}(\theta)}{\Omega(\theta,\theta')}$. The second largest eigenvalue $\lambda_2^{R_B}(\theta, \theta')$ of R_B and second smallest eigenvalue $\lambda_2^{R_A}(\theta, \theta')$ of R_A then satisfy $\lambda_2^{R_B}(\theta, \theta') = 1 - \frac{\lambda_2^{R_A}(\theta)}{\Omega(\theta,\theta')}$. From assumptions 1 and 3, and the lemma's assumptions, $1 - \lambda_2^{R_B}(\theta, \theta') = \frac{\lambda_2^{R_A}(\theta)}{\Omega(\theta,\theta')} \geq \frac{\lambda_2^{R_A}(\theta)}{(K_0+1)\Omega(\theta)} \geq \frac{\mu}{K_0+1}$. Also, since $(1 + \frac{1}{K_0})k_1 \geq 2$,

$$\Omega(\theta, \theta') = \Omega(\theta) + \Omega(\theta') \geq (1 + \frac{1}{K_0})\Omega(\theta) > (1 + \frac{1}{K_0})k_1 \max_s A_s(\theta) \geq 2 \max_s A_s(\theta).$$

So for any state s, the diagonal element $B_s(\theta, \theta') = 1 - \frac{A_s(\theta)}{\Omega(\theta,\theta')} > \frac{1}{2}$. From Fill (1991), this diagonal property and the bound on $1 - \lambda_2^{R_B}(\theta, \theta')$ give the result. □

For Peer Review Only

Our overall proof strategy is to show that on a set with $\|\theta\|$ and $|W|$ large enough, the conditions of Lemma 2 hold with high probability. Lemma 2 then will imply that the distribution over states for the continuous-time MJP and its discrete-time counterpart embedded in W will be brought arbitrarily close to π_θ (and thus to each other), allowing our sampler to inherit mixing properties of the ideal sampler. We will exploit the boundedness of the complement of this set to establish a 'small-set condition' where the MCMC algorithm forgets its state with some probability. These two conditions will be sufficient for geometric ergodicity. The next assumption states these small-set conditions for the ideal sampler.

Assumption 4. *For the ideal sampler with transition probability $p_I(\theta'|\theta)$:*
i) for each M, for the set $B_M = \{\theta : \Omega(\theta) \leq M\}$, there exists a probability measure ϕ and a constant $\kappa_1 > 0$ s.t. $\alpha_I(\theta, \theta'; X)q(\theta'|\theta) \geq \kappa_1\phi(\theta')$ for $\theta \in B_M$. Thus B_M is a 1-small set.
ii) for M large enough, $\exists \rho < 1$ s. t. $\int \Omega(\nu)p_I(\nu|\theta)d\nu \leq (1-\rho)\Omega(\theta) + L_I$, $\forall \theta \notin B_M$.

These two conditions are standard small-set and drift conditions necessary for the ideal sampler to satisfy geometric ergodicity. The first implies that for θ in B_M, the ideal sampler 'forgets' its current location with probability κ_1. The second condition ensures that for θ outside this set, the ideal sampler drifts towards B_M. These two conditions together imply geometric mixing with rate equal or faster than κ_1 (Meyn and Tweedie, 2009). Observe that we have used $\Omega(\theta)$ as the so-called Lyapunov-Foster function to define the drift condition for the ideal sampler. This is the most natural choice, though our proof can be tailored to different choices. Similarly, we could easily allow B_M to be an n-small set for any $n \geq 1$ (so the ideal sampler needs n steps before it can forget its current value in B_M); we restrict ourselves to the 1-small case for clarity.

Assumption 5. $\exists\ u > \ell > 0$ *s.t.* $\prod P(X|s_o, \theta) \in [\ell, u]$ *for any state s_o and θ.*

This assumption follows Miasojedow and Niemiro (2017), and holds if θ does not include parameters of the observation process (or if so, the likelihood is finite and nonzero for all settings of θ). We can relax this assumption, though this will introduce technicalities unrelated to our focus, which is on complications in parameter inference arising from the continuous-time dynamics, rather than the observation process.

For Peer Review Only

Assumption 6. *Given the proposal density* $q(\nu|\theta)$, $\exists \eta_0 > 0, \theta_2 > 0$ *such that for* θ *satisfying* $\|\theta\| > \theta_2$, $\int_\Theta \Omega(\nu)^2 q(\nu|\theta) d\nu \leq \eta_0 \Omega(\theta)^2$.

This mild requirement can be satisfied by choosing a proposal distribution q that does not attempt to explore large θ's too aggressively. The next corollary follows from a simple application of the Cauchy-Schwarz inequality, see the supplement for the proof.

Corollary 1. *Given the proposal density* $q(\nu|\theta)$, $\exists \eta_1 > 0, \theta_2 > 0$ *such that for* θ *satisfying* $\|\theta\| > \theta_2$, $\int_\Theta \Omega(\nu) q(\nu|\theta) d\nu \leq \eta_1 \Omega(\theta)$.

We need two further assumptions on $q(\theta'|\theta)$. There are satisfied in our experiments.

Assumption 7. *For any* $\epsilon > 0$, *there exist finite* M_ϵ, $\theta_{3,\epsilon}$ *such that for* θ *satisfying* $\|\theta\| > \theta_{3,\epsilon}$, *the condition* $q(\{\theta' : \frac{p(\theta')q(\theta|\theta')}{p(\theta)q(\theta'|\theta)} \leq M_\epsilon\}|\theta) > 1 - \epsilon$ *holds.*

This holds, when e.g. $p(\theta)$ is a gamma distribution, and $q(\theta'|\theta)$ is Gaussian.

Assumption 8. *For any* $\epsilon > 0$ *and* $K > 1$, *there exists* $\theta^K_{4,\epsilon}$ *such that for* θ *satisfying* $\|\theta\| > \theta^K_{4,\epsilon}$, *the condition* $q(\{\theta' : \frac{\Omega(\theta')}{\Omega(\theta)} \in \left[\frac{1}{K}, K\right]\}|\theta) > 1 - \epsilon$ *holds.*

This holds when e.g. $q(\theta'|\theta)$ is a centered on θ and has finite variance.

Theorem 3. *Under the above assumptions, our symmetrized auxiliary variable MCMC sampler in algorithm 4 is geometrically ergodic.*

Proof. This theorem follows from two lemmas we will prove. Lemma 5 shows there exist small sets $\{(W, \theta, \vartheta) : \lambda_1 |W| + \Omega(\theta) < M\}$ for $\lambda_1, M > 0$, within which our sampler forgets its current state with some positive probability. Lemma 8 shows that for appropriate (λ_1, M), our sampler drifts towards this set whenever outside. Together, these two results imply geometric ergodicity (Meyn and Tweedie, 2009, Theorems 15.0.1 and Lemma 15.2.8). If $\sup_\theta \Omega(\theta) < \infty$, we just need the small set $\{(W, \theta, \vartheta : |W| < M\}$ for some M. □

For easier comparison with the ideal sampler, we begin an MCMC iteration from step 5 in Algorithm 4. Thus, our sampler operates on (θ, ϑ, W), with θ the current parameter, ϑ the auxiliary variable, and W the Poisson grid. An MCMC iteration updates this by (a) sampling states V with a backward pass, (b) discarding ϑ and self-transition times,

For Peer Review Only

(c) sampling ν from $q(\nu|\theta)$, (d) sampling U' given (θ, ν, S, T), setting $W' = T \cup U'$, and discarding S, (e) proposing to swap (θ, ν) and then (f) accepting or rejecting with a forward pass. On acceptance, $\theta' = \nu$ and $\vartheta' = \theta$, and on rejection, $\theta' = \theta$ and $\vartheta' = \nu$, so that the MCMC state at the end of the iteration is $(\theta', \vartheta', W')$. We write $(\theta'', \vartheta'', W'')$ for the MCMC state after two iterations. Recall that step (a) actually assigns states V to W. T are the elements of W where V changes value, and S are the corresponding elements of V. The remaining elements U are the elements of W corresponding to self-transitions. For reference, we repeat some of our notation in the supplementary material.

We first bound self-transition probabilities of the embedded Markov chain from 0:

Proposition 4. *The posterior probability that the embedded Markov chain makes a self-transition,* $P(V_i = V_{i+1}|W, X, \theta, \vartheta) \geq \delta_1 > 0$, *for any* θ, ϑ, W.

The proof (in the supplement) exploits the bounded likelihood from assumption 5. A simple by-product of the proof is the following corollary:

Corollary 2. $P(V_{i+1} = s|V_s = s, W, X, \theta, \vartheta) \geq \delta_1 > 0$, *for any* θ, ϑ, W, s.

Lemma 5. *For all* $M, h > 0$, *the set* $B_{h,M} = \{(W, \theta, \vartheta) : |W| \leq h, \theta \in B_M\}$ *is a 2-small set under our proposed sampler. Thus, for all* (W, θ, ϑ) *in* $B_{h,M}$, *the two-step transition probability satisfies* $P(W'', \theta'', \vartheta''|W, \theta, \vartheta) \geq \rho_1 \phi_1(W'', \theta'', \vartheta'')$ *for a constant* ρ_1 *and a probability measure* ϕ_1 *independent of the initial state.*

Proof. Recall the definition of B_M, and of an n-small set from Assumption 4. The 1-step transition probability of our MCMC algorithm consists of two terms, corresponding to the proposed parameter being accepted and rejected. Discarding the latter, we have

$$P(W', \theta', \vartheta'|W, \theta, \vartheta, X) \geq q(\theta'|\theta)\delta_\theta(\vartheta')\alpha(\theta, \theta', W'; X) \sum_{S,T} \left[P(S, T|W, \theta, \vartheta, X)P(W'|S, T, \theta, \theta')\right].$$

This follows from steps (c) to (e) in the reordered algorithm. The summation is over all (S, T) values produced by the backward pass (which are then discarded after sampling W'). We have used the fact that given (S, T), $P(W'|S, T, \theta, \theta', X)$ is independent of X.

For Peer Review Only

We bound the summation over (S,T) by considering only terms with S constant. When this constant is state s^*, we write this as $(S=[s^*],T=\emptyset)$. This corresponds to $|W|$ self-transitions after starting state $S_0=s^*$. Then the first term in the square brackets becomes

$$P(S=[s^*],T=\emptyset|W,\theta,\vartheta,X)=P(S_0=s^*|X,W,\theta,\vartheta)\prod_{i=0}^{|W|-1}P(V_{i+1}=s^*|V_i=s^*,X,W,\theta,\vartheta)$$
$$\geq P(S_0=s^*|X,W,\theta,\vartheta)\delta_1^{|W|} \qquad \text{(from Corollary 2).}$$

With $S(t)$ fixed at s^*, W' is distributed as a Poisson process with rate $\Omega(\theta')+\Omega(\theta)-A_{s^*}(\theta)$. Write $\text{PoissProc}(W'|R(t))$ for the probability of W' under a rate-$R(t)$ Poisson process on $[0,t_{end}]$, so that $P(W'|S=[s^*],T=\emptyset,\theta',\theta)=\text{PoissProc}(W'|\Omega(\theta')+\Omega(\theta)-A_{s^*}(\theta))$. Then, from the Poisson superposition theorem, writing $2^{W'}$ for the power set of W', we have

$$P(W'|S=[s^*],T=\emptyset,\theta',\theta)=\sum_{Z\in 2^{W'}}\text{PoissProc}\left(Z|\Omega(\theta')\right)\text{PoissProc}\left(W'\backslash Z|\Omega(\theta)-A_{s^*}(\theta)\right)$$
$$\geq \text{PoissProc}(W'|\Omega(\theta'))\text{PoissProc}(\emptyset|\Omega(\theta)-A_{s^*}(\theta))$$
$$\geq \text{PoissProc}(W'|\Omega(\theta'))\text{PoissProc}(\emptyset|\Omega(\theta))$$
$$\geq \text{PoissProc}(W'|\Omega(\theta'))\exp(-Mt_{end}), \quad \text{since for } \theta\in B_M,\ \Omega(\theta)\leq M.$$

Thus we have

$$\sum_{S,T}P(S,T,W'|W,\theta,\vartheta,X)\geq\sum_{s^*}P(S=[s^*],T=\emptyset|W,\theta,\vartheta,X)P(W'|S=[s^*],T=\emptyset,\theta',\theta)$$
$$\geq \delta_1^{|W|}\exp(-Mt_{end})\text{PoissProc}(W'|\Omega(\theta')). \tag{6}$$

Next, using assumption 5,

$$\alpha(\theta,\theta',W';X)=1\wedge\frac{P(X|W',\theta',\theta)/P(X|\theta')}{P(X|W',\theta,\theta')/P(X|\theta)}\cdot\frac{P(X|\theta')q(\theta|\theta')p(\theta')}{P(X|\theta)q(\theta'|\theta)p(\theta)}$$
$$\geq 1\wedge\frac{\ell^2}{u^2}\cdot\frac{P(X|\theta')q(\theta|\theta')p(\theta')}{P(X|\theta)q(\theta'|\theta)p(\theta)}\geq\alpha_I(\theta,\theta';X)\frac{\ell^2}{u^2}. \tag{7}$$

Inside $B_{h,M}$, $|W|\leq h$, and by assumption 4, $q(\theta'|\theta)\alpha_I(\theta,\theta';X)\geq\kappa_1\phi(\theta')$. Then the three inequalities above let us simplify the equation at the start of the proof:

$$P(W',\theta',\vartheta'|W,\theta,\vartheta)\geq\frac{\ell^2}{u^2}\delta_1^h\exp(-Mt_{end})\delta_\theta(\vartheta')\kappa_1\text{PoissProc}(W'|\Omega(\theta')\phi(\theta')$$
$$\stackrel{\text{def}}{=}\rho_1\delta_\theta(\vartheta')\text{PoissProc}(W'|\Omega(\theta')\phi(\theta').$$

For Peer Review Only

Write $F_{Poiss(a)}$ for the CDF of a rate-a Poisson. The two-step transition satisfies

$$\begin{aligned}
P(W'', \theta'', \vartheta''|W, \theta, \vartheta) &\geq \int_{B_{h,M}} P(W'', \theta'', \vartheta''|W', \theta', \vartheta')P(W', \theta', \vartheta'|W, \theta, \vartheta)\mathrm{d}W'\mathrm{d}\theta'\mathrm{d}\vartheta' \\
&\geq \int_{B_{h,M}} \rho_1 \delta_{\theta'}(\vartheta'')\mathrm{PoissProc}(W''|\Omega(\theta''))\phi(\theta'') \\
&\qquad \rho_1 \delta_{\theta}(\vartheta')\mathrm{PoissProc}(W'|\Omega(\theta'))\phi(\theta')\mathrm{d}W'\mathrm{d}\theta'\mathrm{d}\vartheta' \\
&\geq \rho_1^2 \phi(\theta'')\mathrm{PoissProc}(W''|\Omega(\theta'')) \int_{B_{h,M}} \delta_{\theta'}(\vartheta'')F_{Poiss(\Omega(\theta'))}(h)\phi(\theta')\mathrm{d}\theta' \\
&\geq \rho_1^2 \mathrm{PoissProc}(W''|\Omega(\theta''))\phi(\theta'')\phi(\vartheta'')F_{Poiss(\Omega(\vartheta''))}(h)\delta_{B_{h,M}}(\vartheta'') \\
&\geq \rho_1^2 \mathrm{PoissProc}(W''|\Omega(\theta''))\phi(\theta'')\phi(\vartheta'')\delta_{B_{h,M}}(\vartheta'')\exp(-\Omega(\vartheta'')) \qquad (8)
\end{aligned}$$

The last line uses $F_{Poiss(a)}(h) \geq F_{Poiss(a)}(0) = \exp(-a)\ \forall a$, and gives our result, with $\phi_1(W'', \theta'', \vartheta'') \propto \mathrm{PoissProc}(W''|\Omega(\theta''))\phi(\theta'')\phi(\vartheta'')\delta_{B_{h,M}}(\vartheta'')\exp(-\Omega(\vartheta''))$. □

We have established the small set condition: for a point inside $B_{h,M}$ our sampler forgets its state with nonzero probability, sampling a new state from $\phi_1(\cdot)$. We next establish a drift condition, showing that outside $B_{h,M}$, the algorithm drifts back towards it (Lemma 8). We first establish a result needed when $\max_s |A_s(\theta)|$ is unbounded as θ increases. This states that the acceptance probabilities of our sampler and the ideal sampler can be brought arbitrarily close outside a small set, so long as $\Omega(\theta)$ and $\Omega(\theta')$ are sufficiently close.

Lemma 6. *Suppose* $\frac{1}{K_0} \leq \frac{\Omega(\theta)}{\Omega(\theta')} \leq K_0$, *for* K_0 *satisfying* $(1 + \frac{1}{K_0})k_1 \geq 2$ *(k_1 is from Assumption 1). Write* $|W^{\downarrow}|$ *for the minimum number of elements of grid* W *between any successive pairs of observations. For any* $\epsilon > 0$, *there exist* $w_\epsilon^{K_0}, \theta_{5,\epsilon}^{K_0} > 0$ *such that* $|P(X|W, \theta, \theta') - P(X|\theta)| < \epsilon$ *for any* (W, θ) *with* $|W^{\downarrow}| > w_\epsilon^{K_0}$ *and* $\|\theta\| > \theta_{5,\epsilon}^{K_0}$.

Proof. From lemma 2, for all θ, θ' satisfying the lemma's assumptions, the Markov chain with transition matrix $B(\theta, \theta')$ converges geometrically to stationarity distribution π_θ at a rate uniformly bounded away from 0. By setting $|W^{\downarrow}|$ large enough, for all such (θ, θ') and for any initial state, the Markov chain would have mixed beween each pair of observations, with distribution over states returning arbitrarily close to π_θ.

Write W_X for the indices of the grid W containing observations, and V_X for the Markov chain state at these times (illustrated in Section 11.1 in the supplementary material).

For Peer Review Only

Let $P_B(V_X|W,\theta,\theta')$ be the probability distribution over V_X under the Markov chain with transition matrix B given W and $P_{st}(V_X|\theta)$ be the probability of V_X sampled independently under the stationary distribution. Let $P(X|W,\theta,\theta')$ be the marginal probability of the observations X under that Markov chain $B(\theta,\theta')$ given W. Dropping W and θ' from notation, $P(X|\theta)$ is the probability of the observations under the rate-$A(\theta)$ MJP.

From the first paragraph, for $|W^{\downarrow}| > w_0$ for large enough w_0, $P_B(V_X|W,\theta,\theta')$ and $P_{st}(V_X|W,\theta)$ can be brought ϵ' close. Then for any W with $|W^{\downarrow}| > w_0$, we have

$$|P(X|W,\theta,\theta') - P_{st}(X|\theta)| = |\sum_{V_X} P(X|V_X,\theta)[P_B(V_X|W,\theta,\theta') - P_{st}(V_X|\theta)]|$$
$$\leq \sum_{V_X} P(X|V_X,\theta)|P_B(V_X|W,\theta,\theta') - P_{st}(V_X|\theta)| \leq \epsilon'',$$

using $P(X|V_X,\theta) \leq u$ (Assumption 5), and $\sum_{V_X} |P_B(V_X|W,\theta,\theta') - P_{st}(V_X|\theta)| < \epsilon$. For large θ, we prove a similar result in the continuous case by uniformization. For any θ',

$$P(X|\theta) = \int \mathrm{d}W P(X|W,\theta,\theta')\mathrm{PoissProc}(W|\Omega(\theta)+\Omega(\theta')).$$

We split this integral into two parts, one over the set $\{|W^{\downarrow}| > w_0\}$, and the second over its complement. On the former, for w_0 large enough, $|P(X|W,\theta,\theta') - P_{st}(X|\theta)| \leq \epsilon''$. For θ large enough, $\{|W^{\downarrow}| > w_0\}$ occurs with arbitrarily high probability for any θ'. Since the likelihood is bounded, the integral over the second set can be made arbitrarily small (say, ϵ'' again). Finally, from the triangle inequality,

$$|P(X|\theta) - P(X|W,\theta,\theta')| \leq |P(X|\theta) - P_{st}(X|\theta)| + |P_{st}(X|\theta) - P(X|W,\theta,\theta')|$$
$$\leq (\epsilon'' + \epsilon'') + \epsilon'' \stackrel{\text{def}}{=} \epsilon.$$

□

The previous lemma bounds the difference in probability of observations under the discrete-time and continuous-time processes for θ and $|W|$ large enough. The next result uses this to bound with high probability the different in acceptance probabilities of the ideal sampler, and our proposed sampler with a grid W. See the supplement for the proof.

Proposition 7. *Let (W,θ,ϑ) be the current state of the sampler. Then, for any ϵ, there exists $\theta_\epsilon > 0$ as well as a set $E_\epsilon \subseteq \{(W',\theta') : |\alpha_I(\theta,\theta';X) - \alpha(\theta,\theta';W',X)| \leq \epsilon\}$, such that for θ satisfying $\|\theta\| > \theta_\epsilon$ and any ϑ, we have $P(E_\epsilon|W,\theta,\vartheta) > 1-\epsilon$.*

For Peer Review Only

Lemma 8. *(drift condition) There exist* $\delta_2 \in (0,1), \lambda_1 > 0$ *and* $L > 0$ *such that* $\mathbb{E}\left[\lambda_1|W'| + \Omega(\theta')|W,\theta,\vartheta,X\right] \leq (1-\delta_2)\left(\lambda_1|W| + \Omega(\theta)\right) + L.$

Proof. Since $W' = T \cup U'$, we consider $\mathbb{E}[|T||W,\theta,\vartheta,X]$ and $\mathbb{E}[|U'||W,\theta,\vartheta,X]$ separately. An upper bound of $\mathbb{E}[|T||W,\theta,\vartheta,X]$ can be derived directly from proposition 4:

$$\mathbb{E}[|T||W,\theta,\vartheta,X] = \mathbb{E}[\sum_{i=0}^{|W|-1} \mathbb{I}_{\{V_{i+1}\neq V_i\}}|W,\theta,\vartheta,X] \leq \sum_{i=0}^{|W|-1}(1-\delta_1) = |W|(1-\delta_1).$$

By corollary 1, there exist η_1, θ_2 such that for $\|\theta\| > \theta_2$, $\int \Omega(\nu)q(\nu|\theta)d\nu \leq \eta_1\Omega(\theta)$. Then,

$$\begin{aligned}\mathbb{E}[|U'||W,\theta,\vartheta,X] &= \mathbb{E}_{S,T,\nu}\mathbb{E}[|U'||S,T,W,\theta,\vartheta,\nu,X] = \mathbb{E}_{S,T,\nu}\mathbb{E}[|U'||S,T,W,\theta,\nu] \\ &\leq \mathbb{E}_{S,T,\nu}\left[t_{end}\Omega(\theta,\nu)\right] = t_{end}\int \Omega(\theta,\nu)q(\nu|\theta)\mathrm{d}\nu \\ &= t_{end}\left[\left(\Omega(\theta) + \int_\Theta \Omega(\nu)q(\nu|\theta)\mathrm{d}\nu\right)\right] \leq t_{end}(\eta_1+1)\Omega(\theta).\end{aligned}$$

To bound $\mathbb{E}\left[\Omega(\theta')|W,\theta,\vartheta,X\right]$, consider the transition probability over (W',θ'):

$$\begin{aligned}P(\mathrm{d}W',\mathrm{d}\theta'|W,\theta,\vartheta) = \mathrm{d}\theta'\mathrm{d}W'\Bigg[&q(\theta'|\theta)\sum_{S,T}P(S,T|W,\theta,\vartheta,X)P(W'|S,T,\theta,\theta')\alpha(\theta,\theta';W',X) \\ &+ \int q(\nu|\theta)\sum_{S,T}P(S,T|W,\theta,\vartheta,X)P(W'|S,T,\theta,\nu)(1-\alpha(\theta,\nu;W',X))\mathrm{d}\nu\delta_\theta(\theta')\Bigg].\end{aligned}$$

With $P(W'|W,\theta,\vartheta,\theta',X) = \sum_{S,T}P(S,T|W,\theta,\vartheta,X)P(W'|S,T,\theta,\theta')$, integrate out W':

$$\begin{aligned}P(\mathrm{d}\theta'|W,\theta,\vartheta) = \mathrm{d}\theta'\int \mathrm{d}W'\Bigg[&q(\theta'|\theta)P(W'|W,\theta,\vartheta,\theta',X)\alpha(\theta,\theta';W',X) + \\ &\int q(\nu|\theta)P(W'|W,\theta,\vartheta,\nu,X)(1-\alpha(\theta,\nu;W',X))d\nu\delta_\theta(\theta')\Bigg]\end{aligned}$$

Let $\int \Omega(\theta')P(\mathrm{d}\theta'|W,\theta,\vartheta) = I_1(W,\theta,\vartheta) + \Omega(\theta)I_2(W,\theta,\vartheta)$, with

$$\begin{aligned}I_1(W,\theta,\vartheta) &= \int \mathrm{d}\theta'\Omega(\theta')q(\theta'|\theta)\int \mathrm{d}W'P(W'|W,\theta,\vartheta,\theta',X)\alpha(\theta,\theta';W',X), \\ I_2(W,\theta,\vartheta) &= \int \mathrm{d}\nu dW'q(\nu|\theta)P(W'|W,\theta,\vartheta,\nu,X)(1-\alpha(\theta,\nu;W',X)).\end{aligned}$$

Consider the second term I_2. From Proposition 7, for any positive ϵ, there exists $\theta_\epsilon > 0$ such that the set E_ϵ (where $|\alpha(\theta,\nu;X,W') - \alpha_I(\theta,\nu;X)| \leq \epsilon$) has probability greater than $1-\epsilon$.

For Peer Review Only

Write I_{2,E_ϵ} for the integral restricted to this set, and I_{2,E_ϵ^c} for that over the complement, so that $I_2 = I_{2,E_\epsilon} + I_{2,E_\epsilon^c}$. Then for $\theta > \theta_\epsilon$,

$$\begin{aligned}
I_{2,E_\epsilon}(W,\theta,\vartheta) &= \int_{E_\epsilon} d\nu dW' q(\nu|\theta) P(W'|W,\theta,\vartheta,\nu,X)(1-\alpha(\theta,\nu;W',X)) \\
&\leq \int_{E_\epsilon} d\nu dW' q(\nu|\theta) P(W'|W,\theta,\vartheta,\nu,X)[1-(\alpha_I(\theta,\nu;X)-\epsilon)] \\
&\leq \int d\nu dW' q(\nu|\theta) P(W'|W,\theta,\vartheta,\nu,X)[1-(\alpha_I(\theta,\nu;X)-\epsilon)] \\
&\leq (1+\epsilon) - \int q(\nu|\theta)\alpha_I(\theta,\nu;X)d\nu, \quad \text{and} \\
I_{2,E_\epsilon^c}(W,\theta,\vartheta) &= \int_{E_\epsilon^c} d\nu dW' q(\nu|\theta) P(W'|W,\theta,\vartheta,\nu,X)(1-\alpha(\theta,\nu;W',X)) \\
&\leq \int_{E_\epsilon^c} d\nu dW' q(\nu|\theta) P(W'|W,\theta,\vartheta,\nu,X) \leq \epsilon.
\end{aligned}$$

We similarly divide the integral I_1 into two parts, I_{1,E_ϵ} (over E_ϵ) and I_{1,E_ϵ^c} (over its complement E_ϵ^c). For $\|\theta\|$ large enough, we can bound the acceptance probability by $\alpha_I(\theta,\theta';X)+\epsilon$ on the set E_ϵ, and by corollary 1, we get

$$I_{1,E_\epsilon} \leq \int_{E_\epsilon} \Omega(\theta')q(\theta'|\theta)(\alpha_I(\theta,\theta';X)+\epsilon)d\theta' \leq \int \Omega(\theta')q(\theta'|\theta)\alpha_I(\theta,\theta';X)d\theta' + \eta_1\epsilon\Omega(\theta).$$

For I_{1,E_ϵ^c}, from assumption 6, we have $\int_\Theta \Omega(\nu)^2 q(\nu|\theta)d\nu \leq \eta_0\Omega(\theta)^2$ for $\|\theta\| > \theta_2$. So, by Cauchy-Schwarz inequality and bounding the acceptance probability by one, we have

$$\begin{aligned}
\left(I_{1,E_\epsilon^c}\right)^2 &\leq \int_{E_\epsilon^c} q(\theta'|\theta)P(W'|W,\theta,\vartheta,\theta',X)d\theta' dW' \int_{E_\epsilon^c} \Omega(\theta')^2 q(\theta'|\theta)P(W'|W,\theta,\vartheta,\theta',X)d\theta' dW' \\
&\leq \epsilon \int \Omega(\theta')^2 q(\theta'|\theta)d\theta' \leq \epsilon\eta_0\Omega(\theta)^2,
\end{aligned}$$

giving $I_{1,E_\epsilon^c} \leq \sqrt{\epsilon\eta_0}\Omega(\theta)$. Putting these four results together, for θ satisfying $\|\theta\| > \max(\theta_2,\theta_\epsilon,M)$ (where M is from Assumption 4 on the ideal sampler), we have

$$\begin{aligned}
\int \Omega(\theta')P(d\theta'|W,\theta,\vartheta) &\leq \int \Omega(\theta')q(\theta'|\theta)\alpha_I(\theta,\theta'|X)d\theta' + \Omega(\theta)\int q(\nu|\theta)(1-\alpha_I(\theta,\nu|X))d\nu + \\
&\sqrt{\eta_0\epsilon}\Omega(\theta) + \eta_1\epsilon\Omega(\theta) + 2\epsilon\Omega(\theta) \\
&\leq (1-\rho)\Omega(\theta) + (\sqrt{\eta_0\epsilon} + \eta_1\epsilon + 2\epsilon)\Omega(\theta) + L_I, \quad \text{giving}
\end{aligned}$$

For Peer Review Only

$$\begin{aligned}\mathbb{E}[\lambda_1|W'| + \Omega(\theta')|W, \theta, \vartheta, X] &\leq \lambda_1(1-\delta_1)|W| + \lambda_1 t_{end}(1+\eta_1)\Omega(\theta) + \\ &(1-\rho)\Omega(\theta) + (\sqrt{\eta_0}\sqrt{\epsilon} + \eta_1\epsilon + 2\epsilon)\Omega(\theta) + L_I \\ &= (1-\delta_1)\lambda_1|W| + [1 - (\rho - \lambda_1 t_{end}(1+\eta_1) - (2+\eta_1)\epsilon - \sqrt{\eta_0\epsilon})]\Omega(\theta) + L_I \\ &\stackrel{\text{def}}{=} (1-\delta_1)\lambda_1|W| + (1-\delta_2)\Omega(\theta) + L_I\end{aligned}$$

For (λ_1, ϵ) small enough, $\delta_2 \in (0,1)$, and $\delta = \min(\delta_1, \delta_2)$ gives the drift condition. □

9 Conclusion

We have proposed a novel Metropolis-Hastings algorithm for parameter inference in Markov jump processes. We use uniformization to update the MJP parameters with state-values marginalized out, though still conditioning on a random Poisson grid. The distribution of this grid depends on the MJP parameters, significantly slowing down MCMC mixing. We propose a simple symmetrization scheme to get around this dependency. In our experiments, we demonstrate the usefulness of this scheme, which outperforms a number of competing baselines. We also derive conditions under which our sampler inherits geometric ergodicity properties of an ideal MCMC sampler.

There are a number of interesting directions for future research. Our focus was on Metropolis-Hastings algorithms for typical settings, where the parameters are low dimensional. It is interesting to investigate how our ideas extend to schemes like Hamiltonian Monte Carlo (Neal, 2010) suited for higher-dimensional settings. Another direction is to develop and study similar schemes for more complicated hierarchical models like mixtures of MJPs or coupled MJPs. While we focused only on Markov jump processes, it is also of interest to study similar ideas for algorithms for more general processes (Rao and Teh, 2012). It is also important to investigate how similar ideas apply to deterministic algorithms like variational Bayes (Opper and Sanguinetti, 2007; Pan et al., 2017). From a theoretical viewpoint, our proof required the uniformization rate to satisfy $\Omega(\theta) \geq k_1 \max_s A_s(\theta) + k_0$ for $k_1 > 1$. We believe our result still holds for $k_1 = 1$, and for completeness, it would be interesting to prove this.

For Peer Review Only

10 Supplementary material

Appendix This file includes a summary of notation used in the main text, proofs not included in the main text, details of the naïve and particle MCMC algorithms, as well as experimental results not included in the main text. [`Appendix_ZhangRao.pdf`].

Python code This includes code implementing the symmetrized MH algorithm, as well as the E Coli dataset. `README.txt` includes instructions. The github repository `https://github.com/varao/ZhangRao_JCGS_code` also contains the code. [`Code_ZhangRao.tar.gz`].

11 Acknowledgements

We thank the anonymous reviewers whose suggestions helped to significantly improve this manuscript. We acknowledge the National Science Foundation for funding under grants RI/1816499 and DMS/1812197.

References

Andrieu, C., Doucet, A. and Holenstein., R. (2010), 'Particle Markov chain Monte Carlo methods.', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(3), 269–342.

Andrieu, C. and Roberts, G. O. (2009), 'The pseudo-marginal approach for efficient Monte Carlo computations.', *Annals of Statistics* **37**(2), 697–725.

Breuer, L. (2003), *From Markov jump processes to spatial queues*, Springer.

Carter, C. K. and Kohn, R. (1996), 'Markov chain Monte Carlo in conditionally Gaussian state space models', *Biometrika* **83**(3), 589–601.

Çinlar, E. (1975), *Introduction to Stochastic Processes*, Prentice Hall.

For Peer Review Only

El-Hay, T., Friedman, N. and Kupferman, R. (2008), 'Gibbs Sampling in Factorized Continuous-Time Markov Processes', *in Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pp. 169–178.

Elliott, R. and Osakwe, C. J. (2006), 'Option pricing for pure jump processes with Markov switching compensators', *Finance and Stochastics* **10**, 250–275.

Fearnhead, P. and Sherlock, C. (2006), 'An exact Gibbs sampler for the Markov-modulated Poisson process', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(5), 767–784.

Fill, J. A. (1991), 'Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process', *The Annals of Applied Probability* **1**(1), 62–87.

Frühwirth-Schnatter, S. (1994), 'Data augmentation and dynamic linear models', *Journal of Time Series Analysis* **15**(2), 183–202.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013), *Bayesian Data Analysis*, Chapman and Hall/CRC.

Gillespie, D. T. (1977), 'Exact stochastic simulation of coupled chemical reactions', *The Journal of Physical Chemistry* **81**(25), 2340–2361.

Goldman, N. and Yang, Z. (1994), 'A codon-based model of nucleotide substitution for protein-coding DNA sequences.', *Molecular Biology and Evolution* **11**(5), 725–736.

Gross, D., Shortle, J., Thompson, J. and Harris, C. (2011), *Fundamentals of Queueing Theory*, Wiley Series in Probability and Statistics, Wiley.

Hobolth, A. and Stone, E. (2009), 'Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution.', *The Annals of Applied Statistics* **3**(3), 1204.

Jensen, A. (1953), 'Markoff chains as an aid in the study of Markoff processes', *Skand. Aktuarietiedskr.* **36**, 87–91.

For Peer Review Only

Jukes, T. H. and Cantor, C. R. (1969), *Evolution of Protein Molecules*, Academy Press.

Liu, J. S. (1994), 'The fraction of missing information and convergence rate for data augmentation', *Computing Science and Statistics* pp. 490–490.

Meyn, S. and Tweedie, R. L. (2009), *Markov Chains and Stochastic Stability*, 2nd edn, Cambridge University Press.

Miasojedow, B. and Niemiro, W. (2017), 'Geometric ergodicity of Rao and Teh's algorithm for Markov jump processes and CTBNs', *Electron. J. Statist.* **11**(2), 4629–4648.

Møller, J., Pettitt, A. N., Reeves, R. and Berthelsen, K. K. (2006), 'An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants', *Biometrika* **93**(2), 451–458.

Murray, I., Ghahramani, Z. and MacKay, D. J. C. (2006), MCMC for doubly-intractable distributions, *in Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pp. 359–366.

Neal, R. M. (2004), 'Taking bigger Metropolis steps by dragging fast variables', Technical report, Department of Statistics, University of Toronto.

Neal, R. M. (2010), 'MCMC using Hamiltonian dynamics', *Handbook of Markov Chain Monte Carlo* **54**, 113–162.

Nodelman, U., Shelton, C. and Koller, D. (2002), 'Continuous time Bayesian networks', *in Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pp. 378–387.

Opper, M. and Sanguinetti, G. (2007), 'Variational inference for Markov jump processes', *in Advances in Neural Information Processing Systems 20*, pp. 1105–1112.

Pan, J., Rao, V., Agarwal, P. K. and Gelfand, A. E. (2016), 'Markov-modulated marked Poisson processes for check-in data', *in Proceedings of the 33rd International Conference on Machine Learning*, Vol. 48 of *PMLR*, pp. 2244–2253.

For Peer Review Only

Pan, J., Zhang, B. and Rao, V. (2017), 'Collapsed variational Bayes for Markov jump processes', *in Advances in Neural Information Processing Systems 30*, pp. 3749–3757.

Papaspiliopoulos, O., Roberts, G. O. and Sköld, M. (2007), 'A general framework for the parametrization of hierarchical models', *Statistical Science* pp. 59–73.

Plummer, M., Best, N., Cowles, K. and Vines, K. (2006), 'CODA: Convergence diagnosis and output analysis for MCMC', *R News* **6**(1), 7–11.

Rao, V. and Teh, Y. W. (2011), 'Gaussian process modulated renewal processes', *in Advances in Neural Information Processing Systems 24*, pp. 2474–2482.

Rao, V. and Teh, Y. W. (2012), 'MCMC for continuous-time discrete-state systems', *in Advances in Neural Information Processing Systems 25*, pp. 701–709.

Rao, V. and Teh, Y. W. (2013), 'Fast MCMC sampling for Markov jump processes and extensions.', *Journal of Machine Learning Research* **14**(1).

Scott, S. L. and Smyth, P. (2003), 'The Markov modulated Poisson process and Markov Poisson cascade with applications to web traffic modeling', *Bayesian Statistics* **7**, 1–10.

Xu, J. and Shelton, C. R. (2010), 'Intrusion detection using continuous time Bayesian networks', *Journal of Artificial Intelligence Research* **39**, 745–774.

Yu, Y. and Meng, X.-L. (2011), 'To center or not to center: That is not the question—an Ancillarity–Sufficiency Interweaving Strategy (ASIS) for boosting MCMC efficiency', *Journal of Computational and Graphical Statistics* **20**(3), 531–570.

For Peer Review Only