The Bouncy Particle Sampler: A Non-Reversible Rejection-Free Markov Chain Monte Carlo Method

Alexandre Bouchard-Côté*, Sebastian J. Vollmer† and Arnaud Doucet†

18th January 2017

*Department of Statistics, University of British Columbia, Canada.

†Department of Statistics, University of Oxford, UK.

1 Direct proof of invariance

Let μ_t be the law of $z(t)$. In the following, we prove invariance by explicitly verifying that the time evolution of the density $\frac{\mathrm{d}\mu_t}{\mathrm{d}t} = 0$ is zero if the initial distribution μ_0 is given by $\rho(z) = \pi(x)\psi(v)$ in Proposition 1. This is achieved by deriving the forward Kolmogorov equation describing the evolution of the marginal density of the stochastic process. For simplicity, we start by presenting the invariance argument when $\lambda^{\mathrm{ref}} = 0$.

Notation and description of the algorithm. We denote a pair of position and velocity by $z = (x, v) \in \mathbb{R}^d \times \mathbb{R}^d$ and we denote translations by $\Phi_t(z) = (\Phi_t^{\mathrm{pos}}(z), \Phi_t^{\mathrm{dir}}(z)) = (x + vt, v)$. The time of the first bounce coincides with the first arrival T_1 of a PP with intensity $\chi(t) = \lambda(\Phi_t(z))$ where:

$$\lambda(z) = \max\{0, \langle \nabla U(x), v\rangle\}. \tag{1}$$

It follows that the probability of having no bounce in the interval $[0, t]$ is given by:

$$\mathrm{No}_t(z) = \exp\left(-\int_0^t \lambda(\Phi_s(z))\mathrm{d}s\right), \tag{2}$$

and the density of the random variable T_1 is given by:

$$\begin{aligned} q(t_1; z) = &\quad \mathbf{1}[t_1 > 0]\tfrac{\mathrm{d}}{\mathrm{d}t_1}(1 - \mathrm{No}_{t_1}(z)) &(3)\\ = &\quad \mathbf{1}[t_1 > 0]\mathrm{No}_{t_1}(z)\lambda(\Phi_{t_1}(z)). &(4)\end{aligned}$$

If a bounce occurs, then the algorithm follows a translation path for time T_1, at which point the velocity is updated using a bounce operation $C(z)$, defined as:

$$C(z) = (x, R(x)v) \tag{5}$$

where

$$R(x)v = v - 2\frac{\langle \nabla U(x), v\rangle \nabla U(x)}{\|\nabla U(x)\|^2}. \tag{6}$$

The algorithm then continues recursively for time $t - T_1$, in the following sense: a second bounce time T_2 is simulated by adding to T_1 a random increment with density $q(\cdot; C \circ \Phi_{t_1}(z))$. If $T_2 > t$, then the output of the algorithm is $\Phi_{t-t_1} \circ C \circ \Phi_{t_1}(z)$, otherwise an additional bounce is simulated,

etc. More generally, given an initial point z and a sequence $\mathbf{t} = (t_1, t_2, \dots)$ of bounce times, the output of the algorithm at time t is given by:

$$\Psi_{\mathbf{t},t}(z) = \begin{cases} \Phi_t(z) & \text{if } t_1 > 0 \text{ or } \mathbf{t} = (\,), \\ \Psi_{\mathbf{t}',t-t_1}(z) \circ C \circ \Phi_{t_1}(z) & \text{otherwise,} \end{cases} \tag{7}$$

where $(\,)$ denotes the empty list and $\mathbf{t}'$ the suffix of $\mathbf{t}$: $\mathbf{t}' = (t_2, t_3, \dots)$. As for the bounce times, they are distributed as follows:

$$T_1 \sim q(\,\cdot\,; z) \tag{8}$$

$$T_i - T_{i-1} | T_{1:i-1} \sim q\Big(\,\cdot\,; \underbrace{\Psi_{T_{1:i-1},T_{i-1}}(z)}_{\text{Pos. after collision } i-1}\Big), \quad i \in \{2, 3, 4, \dots\} \tag{9}$$

where $T_{1:i-1} = (T_1, T_2, \dots, T_{i-1})$.

Decomposition by the number of bounces. Let h denote an arbitrary non-negative measurable test function. We show how to decompose expectations of the form $\mathbb{E}[h(\Psi_{\mathbf{T},t}(z))]$ by the number of bounces in the interval $(0, t)$. To do so, we introduce a function $\#\mathrm{Col}_t(\mathbf{t})$, which returns the number of bounces in the interval $(0, t)$:

$$\#\mathrm{Col}_t(\mathbf{t}) = \min\{n \geq 1 : t_n > t\} - 1. \tag{10}$$

From this, we get the following decomposition:

$$\mathbb{E}[h(\Psi_{\mathbf{T},t}(z))] = \mathbb{E}[h(\Psi_{\mathbf{T},t}(z)) \sum_{n=0}^{\infty} \mathbf{1}[\#\mathrm{Col}_t(\mathbf{T}) = n]] \tag{11}$$

$$= \sum_{n=0}^{\infty} \mathbb{E}[h(\Psi_{\mathbf{T},t}(z)) \mathbf{1}[\#\mathrm{Col}_t(\mathbf{T}) = n]]. \tag{12}$$

On the event that no bounce occurs in the interval $[0, t)$, i.e. $\#\mathrm{Col}_t(\mathbf{T}) = 0$, the function $\Psi_{\mathbf{T},t}(z)$ is equal to $\Phi_t(z)$, therefore:

$$\mathbb{E}[h(\Psi_{\mathbf{T},t}(z)) \mathbf{1}[\#\mathrm{Col}_t(\mathbf{T}) = 0]] = h(\Phi_t(z)) \mathbb{P}(\#\mathrm{Col}_t(\mathbf{T}) = 0) \tag{13}$$

$$= h(\Phi_t(z)) \mathrm{No}_t(z). \tag{14}$$

Indeed, on the event that $n \geq 1$ bounces occur, the random variable $h(\Phi_t(z))$ only depends on a finite dimensional random vector, $(T_1, T_2, \dots, T_n)$, so we can write the expectation as an integral with respect to the density $\widetilde{q}(t_{1:n}; t, z)$ of these variables:

$$\begin{aligned} &\mathbb{E}[h(\Psi_{\mathbf{T},t}(z)) \mathbf{1}[\#\mathrm{Col}_t(\mathbf{T}) = n]] \qquad (15) \\ &= \mathbb{E}\left[h(\Psi_{\mathbf{T},t}(z)) \mathbf{1}[0 < T_1 < \cdots < T_n < t < T_{n+1}]\right] \\ &= \int \cdots \int_{0<t_1<\cdots<t_n<t<t_{n+1}} h(\Psi_{t_{1:n},t}(z)) q(t_1; z) \prod_{i=2}^{n+1} q(t - t_{i-1}; \Psi_{t_{1:i-1},t_{i-1}}(z)) \mathrm{d}t_{1:n+1} \\ &= \int \cdots \int_{0<t_1<\cdots<t_n<t} h(\Psi_{t_{1:n},t}(z)) \widetilde{q}(t_{1:n}; t, z) \mathrm{d}t_{1:n}, \qquad (16) \end{aligned}$$

where:

$$\widetilde{q}(t_{1:n}; t, z) = q(t_1; z) \times \begin{cases} \mathrm{No}_{t-t_1}(\Phi_{t_1}(z)) & \text{if } n = 1 \\ \mathrm{No}_{t-t_n}(\Phi_{t_{1:n},t_n}(z)) & \prod_{i=2}^{n} q\left(t_i - t_{i-1}; \Psi_{t_{1:i-1},t_{i-1}}(z)\right) \text{ if } n \geq 2. \end{cases}$$

To include Equations (14) and (16) under the same notation, we define $t_{1:0}$ to the empty list, $(\,)$, $\widetilde{q}((\,); t, z) = \mathrm{No}_t(z)$, and abuse the integral notation so that for all $n \in \{0, 1, 2, \dots\}$:

$$\mathbb{E}[h(\Psi_{\mathbf{T},t}(z)) \mathbf{1}[\#\mathrm{Col}_t(\mathbf{T}) = n]] = \int \cdots \int_{0<t_1<\cdots<t_n<t} h(\Psi_{t_{1:n},t}(z)) \widetilde{q}(t_{1:n}; t, z) \mathrm{d}t_{1:n}. \tag{17}$$

Marginal density. Let us fix some arbitrary time $t > 0$. We seek a convenient expression for the marginal density at time t, $\mu_t(z)$, given an initial vector $Z \sim \rho$, where ρ is the hypothesized stationary density $\rho(z) = \pi(x)\psi(v)$ on Z. To do so, we look at the expectation of an arbitrary non-negative measurable test function h:

$$\begin{aligned}
\mathbb{E}[h(\Psi_{\mathbf{T},t}(Z))] &= \mathbb{E}\Big[\mathbb{E}[h(\Psi_{\mathbf{T},t}(Z))|Z]\Big] && (18)\\
&= \sum_{n=0}^{\infty} \mathbb{E}\Big[\mathbb{E}[h(\Psi_{\mathbf{T},t}(Z))\mathbf{1}[\#\mathrm{Col}_t(\mathbf{T}) = n]|Z]\Big] && (19)\\
&= \sum_{n=0}^{\infty} \int_{\mathcal{Z}} \rho(z) \int \cdots \int_{0<t_1<\cdots<t_n<t} h(\Psi_{t_{1:n},t}(z))\widetilde{q}(t_{1:n};t,z)\mathrm{d}t_{1:n}\mathrm{d}z && (20)\\
&= \sum_{n=0}^{\infty} \int \cdots \int_{0<t_1<\cdots<t_n<t} \int_{\mathcal{Z}} \rho(z)h(\Psi_{t_{1:n},t}(z))\widetilde{q}(t_{1:n};t,z)\mathrm{d}z\mathrm{d}t_{1:n} && (21)\\
&= \sum_{n=0}^{\infty} \int \cdots \int_{0<t_1<\cdots<t_n<t} \int_{\mathcal{Z}} \rho(\Psi^{-1}_{t_{1:n},t}(z'))h(z')\widetilde{q}(t_{1:n};t,\Psi^{-1}_{t_{1:n},t}(z'))\left|\det D\Psi^{-1}_{t_{1:n},t}\right|\mathrm{d}z'\mathrm{d}t_{1:n}\\
&= \int_{\mathcal{Z}} h(z') \underbrace{\sum_{n=0}^{\infty} \int \cdots \int_{0<t_1<\cdots<t_n<t} \rho(\Psi^{-1}_{t_{1:n},t}(z'))\widetilde{q}(t_{1:n};t,\Psi^{-1}_{t_{1:n},t}(z'))\mathrm{d}t_{1:n}}_{\mu_t(z')}\,\mathrm{d}z'. && (22)
\end{aligned}$$

We used the following in the above derivation successively the law of total expectation, equation (12), equation (18), Tonelli's theorem and the change of variables, $z' = \Psi_{t_{1:n},t}(z)$, justified since for any fixed $0 < t_1 < t_2 < \cdots < t_n < t < t_{n+1}$, $\Psi_{t_{1:n},t}(\cdot)$ is a bijection (being a composition of bijections). Now the absolute value of the determinant is one since $\Psi_{\mathbf{t},t}(z)$ is a composition of unit-Jacobian mappings and, by using Tonelli's theorem again, we obtain that the expression above the brace is necessarily equal to $\mu_t(z')$ since h is arbitrary.

Derivative. Our goal is to show that for all $z' \in \mathcal{Z}$

$$\frac{\mathrm{d}\mu_t(z')}{\mathrm{d}t} = 0.$$

Since the process is time homogeneous, once we have computed the derivative, it is enough to show that it is equal to zero at $t = 0$. To do so, we decompose the computation according to the terms I_n in Equation (22):

$$\begin{aligned}
\mu_t(z') &= \sum_{n=0}^{\infty} I_n(z',t) && (23)\\
I_n(z',t) &= \int \cdots \int_{0<t_1<\cdots<t_n<t} \rho(\Psi^{-1}_{t_{1:n},t}(z'))\widetilde{q}(t_{1:n};t,\Psi^{-1}_{t_{1:n},t}(z'))\mathrm{d}t_{1:n}. && (24)
\end{aligned}$$

The categories of terms in Equation (23) to consider are:

No bounce: $n = 0$, $\Psi_{t_{1:n},t}(z) = \Phi_t(z)$, or,

Exactly one bounce: $n = 1$, $\Psi_{t_{1:n},t}(z) = F_{t,t_1} := \Phi_{t-t_1} \circ C \circ \Phi_{t_1}(z)$ for some $t_1 \in (0,t)$, or,

Two or more bounces: $n \geq 2$, $\Psi_{t_{1:n},t}(z) = \Psi_{t-t_2} \circ C \circ F_{t_2,t_1}(z)$ for some $0 < t_1 < t_2 < t$

In the following, we show that the derivative of the terms in the third category, $n \geq 2$, are all equal to zero, while the derivative of the first two categories cancel each other.

No bounce in the interval. From Equation (14):

$$I_0(z',t) = \rho(\Phi_{-t}(z'))\mathrm{No}_t(\Phi_{-t}(z')). \quad (25)$$

We now compute the derivative at zero of the above expression:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} I_0(z',t)\Big|_{t=0} = \mathrm{No}_0 \quad & (\Phi_0(z'))\ \frac{\mathrm{d}\rho(\Phi_{-t}(z'))}{\mathrm{d}t}\Big|_{t=0} + \\
& \rho(\Phi_0(z'))\ \frac{\mathrm{d}\mathrm{No}_t(\Phi_{-t}(z'))}{\mathrm{d}t}\Big|_{t=0} \qquad (26)
\end{aligned}$$

The first term in the above equation can be simplified as follows:

$$\begin{aligned}
\mathrm{No}_0(\Phi_0(z'))\frac{\mathrm{d}\rho(\Phi_{-t}(z'))}{\mathrm{d}t} &= \frac{\mathrm{d}\rho(\Phi_{-t}(z'))}{\mathrm{d}t} && (27)\\
&= \left\langle \frac{\partial\rho(\Phi_{-t}(z'))}{\partial\Phi^{\mathrm{pos}}_{-t}(z')}, \frac{\mathrm{d}\Phi^{\mathrm{pos}}_{-t}(z')}{\mathrm{d}t} \right\rangle + \\
&\quad \left\langle \frac{\partial\rho(\Phi_{-t}(z'))}{\partial\Phi^{\mathrm{dir}}_{-t}(z')}, \underbrace{\frac{\mathrm{d}\Phi^{\mathrm{dir}}_{-t}(z')}{\mathrm{d}t}}_{=\mathbf{0}} \right\rangle && (28)\\
&= \left\langle \frac{\partial\rho(z)}{\partial x}, -v' \right\rangle && (29)\\
&= \left\langle \frac{\partial}{\partial x}\frac{1}{Z}\exp\left(-U(x)\right)\psi\left(v\right), -v' \right\rangle \\
&= \rho(\Phi_{-t}(z'))\left\langle \nabla U(x), v' \right\rangle, && (30)
\end{aligned}$$

where $x = \Phi^{\mathrm{pos}}_{-t}(z')$. The second term in Equation (26) is equal to:

$$\begin{aligned}
\rho(\Phi_0(z'))\left.\frac{\mathrm{dNo}_t(\Phi_{-t}(z'))}{\mathrm{d}t}\right|_{t=0} &= -\rho(\Phi_0(z'))\mathrm{No}_0(z')\lambda(\Phi_0(z')) && (31)\\
&= -\rho(z')\lambda(z'), && (32)
\end{aligned}$$

using Equation (4). In summary, we have:

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}I_0(z',t)\right|_{t=0} = \rho(z')\left\langle \nabla U(x'), v' \right\rangle - \rho(z')\lambda(z').$$

Exactly one bounce in the interval. From Equation (16), the trajectory consists in a bounce at a time T_1, occurring with density (expressed as before as a function of the final point z') $q(t_1; F^{-1}_{t,t_1}(z'))$, followed by no bounce in the interval $(T_1, t]$, an event of probability:

$$\begin{aligned}
\mathrm{No}_{t-t_1}(C\circ\Phi_{t_1}(z)) &= \mathrm{No}_{t-t_1}(C\circ\Phi_{t_1}\circ F^{-1}_{t,t_1}(z')) && (33)\\
&= \mathrm{No}_{t-t_1}(\Phi_{t_1-t}(z')), && (34)
\end{aligned}$$

where we used that $C^{-1} = C$. This yields:

$$I_1(z',t) = \int_0^t q(t_1; F^{-1}_{t,t_1}(z'))\rho(\Psi^{-1}_{t_1:1,t}(z'))\mathrm{No}_{t-t_1}(\Phi_{t_1-t}(z'))\mathrm{d}t_1. \quad (35)$$

To compute the derivative of the above equation at zero, we use again Leibniz's rule:

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}I_1(z',t)\right|_{t=0} = \rho(C(z'))\lambda(C(z')).$$

Two or more bounces in the interval. For a number of bounce, we get:

$$I_n(z',t) = \int_0^t \Bigg[\underbrace{\int\cdots\int_{t_{2:n}:t_1<t_2\cdots<t_n<t} \rho(\Psi^{-1}_{t_{1:n},t}(z'))\widetilde{q}(t_{1:n}; t, \Psi^{-1}_{t_{1:n},t}(z'))\mathrm{d}t_{2:n}}_{\tilde{I}(t_1,t,z')} \Bigg]\mathrm{d}t_1, \quad (36)$$

and hence, using Leibniz's rule on the integral over t_1:

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}I_n(z',t)\right|_{t=0} = \tilde{I}(0,0,z') = 0. \quad (37)$$

Putting all terms together. Putting everything together, we obtain:

$$\left.\frac{\mathrm{d}\mu_t(z')}{\mathrm{d}t}\right|_{t=0} = \rho(z')\left\langle \nabla U(x'), v' \right\rangle \underbrace{-\rho(z')\lambda(z') + \rho(C(z'))\lambda(C(z'))}. \quad (38)$$

From the expression of $\lambda(\cdot)$, we can rewrite the two terms above the brace as follows:

$$\begin{aligned}
&-\rho(z')\lambda(z')+\rho(C(z'))\lambda(C(z'))\\
=&-\rho(z')\lambda(z')+\rho(z')\lambda(C(z'))\\
=&-\rho(z')\max\{0,\langle\nabla U(x'),v'\rangle\}+\rho(z')\max\{0,\langle\nabla U(x'),R(x')\,v'\rangle\}\\
=&-\rho(z')\max\{0,\langle\nabla U(x'),v'\rangle\}+\rho(z')\max\{0,\langle\nabla U(x'),R(x')\,v'\rangle\}\\
=&-\rho(z')\max\{0,\langle\nabla U(x'),v'\rangle\}+\rho(z')\max\{0,-\langle\nabla U(x'),v'\rangle\}\\
=&-\rho(z')\,\langle\nabla U(x'),v'\rangle\,,
\end{aligned}$$

where we used that $\rho(z')=\rho(C(z'))$, $\langle\nabla U(x'),R(x')\,v'\rangle=-\langle\nabla U(x'),v'\rangle$ and $-\max\{0,f\}+\max\{0,-f\}=-f$ for any function f. Hence we have $\frac{\mathrm{d}\mu_t(z')}{\mathrm{d}t}\Big|_{t=0}=0$, establishing that that the bouncy particle sampler $\lambda^{\mathrm{ref}}=0$ admits ρ as invariant distribution. The invariance for $\lambda^{\mathrm{ref}}>0$ then follows from Lemma 1 given below.

Lemma 1. *Suppose P_t is a continuous time Markov kernel and Q is a discrete time Markov kernel which are both invariant with respect to μ. Suppose we construct for $\lambda^{ref}>0$ a Markov process $\hat{P}_t$ as follows: at the jump times of an independent PP with intensity λ^{ref} we make a transition with Q and then continue according to P_t, then $\hat{P}_t$ is also μ-invariant.*

Proof. The transition kernel is given by

$$\begin{aligned}
\hat{P}_t &= e^{-\lambda t}P_t+\int_0^t \mathrm{d}t_1\lambda e^{\lambda t_1}e^{-\lambda(t-t_1)}P_{t-t_1}QP_{t_1}\\
&\quad+\int_0^t \mathrm{d}t_1\int_{t_1}^{t_2}\mathrm{d}t_2\lambda^2 e^{\lambda t_1}e^{\lambda(t_2-t_1)}e^{-\lambda(t-t_2)}P_{t-t_2}QP_{t_2-t_1}QP_{t_1}+\ldots
\end{aligned}$$

Therefore

$$\begin{aligned}
\mu\hat{P}_t &= \mu\left(e^{-\lambda t}+\lambda te^{-\lambda t}+\frac{(\lambda t)^2}{2}e^{-\lambda t}\ldots\right)\\
&= \mu.
\end{aligned}$$

Hence $\hat{P}_t$ is μ-invariant. □

2 Invariance of the local sampler

The generator of the local BPS is given by

$$\begin{aligned}
\mathcal{L}h(z) &= \langle\nabla_x h(x,v),v\rangle\\
&\quad+\sum_{f\in F}\lambda_f(x,v)\{h(x,R_f(x)\,v)-h(x,v)\}\\
&\quad+\lambda^{\mathrm{ref}}\int(h(x,v')-h(x,v))\,\psi(\mathrm{d}v')\,.
\end{aligned}\tag{39}$$

The proof of invariance of the local BPS is very similar to the proof of Propostion 1. We have

$$\begin{aligned}
\int\mathcal{L}h(z)\rho(z)\,\mathrm{d}z &= \int\int\langle\nabla_x h(x,v),v\rangle\,\rho(z)\,\mathrm{d}z &(40)\\
&\quad+\int\int\sum_{f\in F}\lambda_f(x,v)\{h(x,R_f(x)\,v)-h(x,v)\}]\rho(z)\,\mathrm{d}z &(41)\\
&\quad+\lambda^{\mathrm{ref}}\int\int\int(h(x,v')-h(x,v))\,\psi(\mathrm{d}v')\,\rho(z)\,\mathrm{d}z &(42)
\end{aligned}$$

where the term (42) is straightforwardly equal to 0 while, by integration by parts, the term (40) satisfies

$$\int\int \langle \nabla_x h\left(x,v\right),v\rangle \,\rho\left(z\right)\mathrm{d}z = \int\int \langle \nabla U\left(x\right),v\rangle \,h\left(x,v\right)\rho\left(z\right)\mathrm{d}z. \tag{43}$$

as h is bounded. Now a change-of-variables shows that for any $f \in F$

$$\int\int \lambda_f\left(x,v\right) h(x,R_f\left(x\right)v)\rho\left(z\right)\mathrm{d}z = \int\int \lambda\left(x,R_f\left(x\right)v\right) h(x,v)\rho\left(z\right)\mathrm{d}z \tag{44}$$

as $R_f^{-1}\left(x\right)v) = R\left(x\right)v$ and $\left\|R_f\left(x\right)v\right\| = \left\|v\right\|$ implies $\psi\left(R_f\left(x\right)v\right) = \psi\left(v\right)$. So the term (41) satisfies

$$\begin{aligned}
&\int\int \sum_{f\in F} \lambda_f\left(x,v\right)\left\{h(x,R_f\left(x\right)v) - h(z)\right\}]\rho\left(z\right)\mathrm{d}z \\
=\ &\int\int \sum_{f\in F} [\lambda\left(x,R_f\left(x\right)v\right) - \lambda\left(x,v\right)]h(x,v)\rho\left(z\right)\mathrm{d}z \\
=\ &\int\int \sum_{f\in F} [\max\{0,\langle \nabla U_f(x), R\left(x\right)v\rangle\} - \max\{0,\langle \nabla U_f(x),v\rangle\}]h(x,v)\rho\left(z\right)\mathrm{d}z \\
=\ &\int\int \sum_{f\in F} [\max\{0,-\langle \nabla U_f(x),v\rangle\} - \max\{0,\langle \nabla U_f(x),v\rangle\}]h(x,v)\rho\left(z\right)\mathrm{d}z \\
=\ &-\int\int \sum_{f\in F} [\langle \nabla U_f\left(x\right),v\rangle]h(x,v)\rho\left(z\right)\mathrm{d}z \\
=\ &-\int\int \langle \nabla U\left(x\right),v\rangle]h(x,v)\rho\left(z\right)\mathrm{d}z,
\end{aligned} \tag{45}$$

where we have used $\langle \nabla U_f(x), R_f\left(x\right)v\rangle = -\langle \nabla U_f(x),v\rangle$ and $\max\{0,-f\} - \max\{0,f\} = -f$ for any f. Hence, summing (43)-(45)-(42), we obtain $\int \mathcal{L}h(z)\rho\left(z\right)\mathrm{d}z = 0$ and the result now follows by [1, Proposition 34.7].

3 Calculations in the isotropic normal case

As we do not use refreshment, it follows from the definition of the collision operator that

$$\begin{aligned}
\left\langle x^{(i)},v^{(i)}\right\rangle &= \left\langle x^{(i)},v^{(i-1)} - \frac{2\left\langle x^{(i)},v^{(i-1)}\right\rangle}{\left\|x^{(i)}\right\|^2}x^{(i)}\right\rangle \\
&= -\left\langle x^{(i)},v^{(i-1)}\right\rangle = -\left\langle x^{(i-1)},v^{(i-1)}\right\rangle - \tau_i \\
&= \begin{cases} -\sqrt{-\log V_i} & \text{if } \left\langle x^{(i-1)},v^{(i-1)}\right\rangle \le 0 \\ -\sqrt{\left\langle x^{(i-1)},v^{(i-1)}\right\rangle^2 - \log V_i} & \text{otherwise} \end{cases},
\end{aligned}$$

and therefore

$$\left\|x^{(i)}\right\|^2 = \begin{cases} \left\|x^{(i-1)}\right\|^2 - \left\langle x^{(i-1)},v^{(i-1)}\right\rangle^2 - \log V_i & \text{if } \left\langle x^{(i-1)},v^{(i-1)}\right\rangle \le 0 \\ \left\|x^{(i-1)}\right\|^2 - \log V_i & \text{otherwise.} \end{cases}.$$

It follows that $\left\langle x^{(j)},v^{(j)}\right\rangle \le 0$ for $j > 0$ if $\left\langle x^{(0)},v^{(0)}\right\rangle \le 0$ so, in this case, we have

$$
\begin{aligned}
\left\|x^{(i)}\right\|^2 &= \left\|x^{(i-1)}\right\|^2 - \left\langle x^{(i-1)}, v^{(i-1)}\right\rangle^2 - \log V_i \\
&= \left\|x^{(i-1)}\right\|^2 + \log V_{i-1} - \log V_i \\
&= \left\|x^{(i-2)}\right\|^2 - \left\langle x^{(i-1)}, v^{(i-1)}\right\rangle^2 - \log V_{i-1} + \log V_{i-1} - \log V_i \\
&\vdots \\
&= \left\|x^{(1)}\right\|^2 - \left\langle x^{(1)}, v^{(1)}\right\rangle^2 - \log V_i
\end{aligned}
$$

In particular for $x^{(0)} = e_1$ and $v^{(0)} = e_2$ with e_i being elements of standard basis of $\mathbb{R}^d$, the norm of the position at all points along the trajectory can never be smaller than 1.

4 Supplementary information on the evolutionary parameters inference experiments

4.1 Model

We consider an over-parameterized generalized time reversible rate matrix [2] with $d = 10$ corresponding to 4 unnormalized stationary parameters $x_1, \ldots, x_4$, and 6 unconstrained substitution parameters $x_{\{i,j\}}$, which are indexed by sets of size 2, i.e. where $i, j \in \{1, 2, 3, 4\}$, $i \neq j$. Off-diagonal entries of Q are obtained via $q_{i,j} = \pi_j \exp\left(x_{\{i,j\}}\right)$, where

$$
\pi_j = \frac{\exp\left(x_j\right)}{\sum_{k=1}^{4} \exp\left(x_k\right)}.
$$

We assign independent standard Gaussian priors on the parameters x_i. We assume that a matrix of aligned nucleotides is provided, where rows are species and columns contains nucleotides believed to come from a shared ancestral nucleotide. Given $x = \left(x_1, \ldots, x_4, x_{\{1,2\}}, \ldots, x_{\{3,4\}}\right)$, and hence Q, the likelihood is a product of conditionally independent continuous time Markov chains over {A, C, G, T}, with "time" replaced by a branching process specified by the phylogenetic tree's topology and branch lengths. The parameter x is unidentifiable, and while this can be addressed by bounded or curved parameterizations, the over-parameterization provides an interesting challenge for sampling methods, which need to cope with the strong induced correlations.

4.2 Baseline

We compare the BPS against a state-of-the-art HMC sampler [3] that uses Bayesian optimization to adapt the the leap-frog stepsize ϵ and trajectory length L of HMC. This sampler was shown in [4] to be comparable or better to other state-of-the-art HMC methods such as NUTS. It also has the advantage of having efficient implementations in several languages. We use the author's Java implementation to compare to our Java implementation of the BPS. Both methods view the objective function as a black box (concretely, a Java interface supporting pointwise evaluation and gradient calculation). In all experiments, we initialize at the mode and use a burn-in of 100 iterations and no thinning. The HMC auto-tuner yielded $\epsilon = 0.39$ and $L = 100$. For our method, we use the global sampler and the global refreshment scheme.

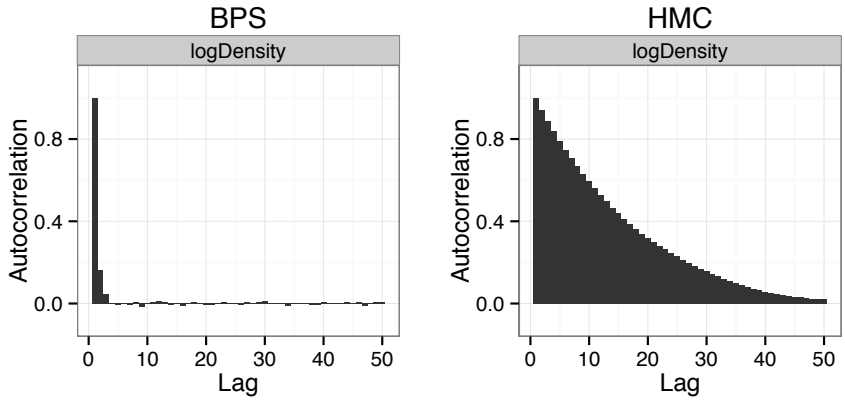

Figure 1: Estimate of the ACF of the log-likelihood statistic for BPS (left) and HMC (right). A similar behavior is observed for the ACF of the other statistics.

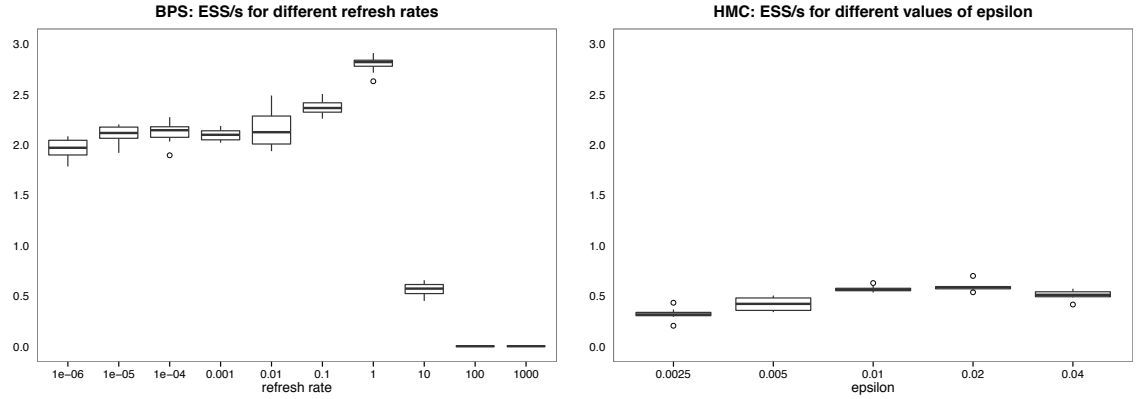

Figure 2: Left: sensitivity of BPS's ESS/s on the log likelihood statistic. Right: sensitivity of HMC's ESS/s on the log likelihood statistic. Each setting is replicated 10 times with different algorithmic random seeds.

4.3 Additional experimental results

To ensure that BPS outperforming HMC does not come from a faulty auto-tuning of HMC parameters, we look at the ESS/s for the log-likelihood statistic when varying the stepsize ϵ. The results in Figure 2(right) show that the value selected by the auto-tuner is indeed reasonable, close to the value 0.02 found by brute force maximization. We repeat the experiments with $\epsilon = 0.02$ and obtain the same conclusions. This shows that the problem is genuinely challenging for HMC.

The BPS algorithm also exhibits sensitivity to λ^{ref}. We analyze this dependency in Figure 2(left). We observe an asymmetric dependency, where values higher than 1 result in a significant drop in performance, as they bring the sampler closer to random walk behavior. Values one or more orders of magnitudes lower than 1 have a lower detrimental effect. However for a range of values of λ^{ref} covering six orders of magnitudes, BPS outperforms HMC at its optimal parameters.

References

[1] M.H.A. Davis. *Markov Models & Optimization*, volume 49. CRC Press, 1993.

[2] S. Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17:56–86, 1986.

[3] Z. Wang, S. Mohamed, and N. de Freitas. Adaptive Hamiltonian and Riemann manifold Monte Carlo. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1462–1470, 2013.

[4] T Zhao, Z. Wang, A. Cumberworth, J. Gsponer, N. de Freitas, and A. Bouchard-Côté. Bayesian analysis of continuous time Markov chains with application to phylogenetic modelling. *Bayesian Analysis*, 2016. To appear.