A Metropolis-class sampler for targets with non-convex support

John Moriarty[1] · Jure Vogrinc[2] · Alessandro Zocca[3]

Abstract

We aim to improve upon the exploration of the general-purpose random walk Metropolis algorithm when the target has non-convex support $A \subset \mathbb{R}^d$, by reusing proposals in A^c which would otherwise be rejected. The algorithm is Metropolis-class and under standard conditions the chain satisfies a strong law of large numbers and central limit theorem. Theoretical and numerical evidence of improved performance relative to random walk Metropolis are provided. Issues of implementation are discussed and numerical examples, including applications to global optimisation and rare event sampling, are presented.

Keywords Markov Chain Monte Carlo · Metropolis-Hastings algorithm · Multimodal target distribution · Multistart method · Global optimisation

Mathematics Subject Classification 65C05 · 62F12 · 60F05 · 60J05 · 65C40 · 90C26 · 65K05

1 Introduction

A key challenge for Markov chain Monte Carlo (MCMC) algorithms is the balance between global "exploration" and local "exploitation". In this paper we present the *skipping sampler*, a general-purpose and easily implemented Metropolis-class algorithm which is capable of improving exploration of targets π with nontrivial support A, by reusing proposals lying outside A. For this to be useful, we make the following standing assumption:

Assumption 1 π is a probability density function on $\mathbb{R}^d$ whose support

$$A = \mathrm{supp}(\pi) := \{x \in \mathbb{R}^d : \pi(x) > 0\}$$

✉ Alessandro Zocca
  a.zocca@vu.nl

  John Moriarty
  j.moriarty@qmul.ac.uk

  Jure Vogrinc
  jure.vogrinc@warwick.ac.uk

[1] School of Mathematics, Queen Mary University of London, London E1 4NS, UK

[2] Department of Statistics, University of Warwick, Coventry CV4 7AL, England

[3] Department of Mathematics, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, Netherlands

satisfies $Leb(A^c) > 0$, where A^c is the complement of A and Leb denotes Lebesgue measure on $\mathbb{R}^d$.

Such targets can arise for example in sampling from the superlevel sets of a density in the hybrid slice sampler Neal (2003), or when sampling from rare events.

Proposals in A^c would be automatically rejected by standard algorithms such as random walk Metropolis (RWM), which exploits only local proposals for the next state of the chain. If a proposal lies in A^c, the skipping sampler uses this information by attempting to cross A^c in a sequence of linear steps, much as a flat stone can jump repeatedly across the surface of water, and offer a relevant proposal. Since this can be seen as a tunnelling effect through the zero-mass region A^c, it is advantageous when A is non-convex and, in particular, disconnected. The resulting Markov chain satisfies a strong law of large numbers and central limit theorem under essentially the same conditions as for RWM, to which we provide theoretical and numerical performance comparisons.

To accelerate global exploration of the state space in MCMC algorithms, several approaches have by now been developed including tempering, Hamiltonian Monte Carlo and piecewise deterministic methods (see Robert et al. (2018) for a recent review). However these methods are best suited to target densities with connected support, since the chain cannot cross regions where the target has zero density. A disconnected support would thus imply reducibility of the chain and its failure to converge to the target.

While RWM can be applied to targets with regions of zero density, its balance between exploration and exploitation can be problematic. If any state in A^c is proposed it is discarded and the chain does not progress. When A is non-convex, in particular, examples may be constructed where exploration is slow even when RWM is well tuned, making the chain sensitive to its initial state. This is illustrated in Fig. 1, where red dots show the trace of a tuned RWM applied to a target with non-convex support, with four different initial states of the chain (the blue traces illustrate the increased exploration achieved by the skipping sampler). One solution is to use knowledge of the target to design a more advanced proposal, such as those reviewed in Robert et al. (2018). However this approach is unavailable if the target density is unknown, or is known but insufficiently regular. In this case, a general-purpose method is instead required.

Theorem 2 establishes that the performance of the skipping sampler is at least as good as that of RWM according to the Peskun ordering. However the strengths of the proposed method lie primarily in applications to difficult low-dimensional problems. Conversely, in high dimensional problems the method generally offers similar performance to RWM. The aim of this paper is to present the method and illustrate its benefits via numerical examples, rather than to study any particular application exhaustively.

Although it is not random walk-based, the skipping sampler is Metropolis class. The symmetry of the skipping proposal can be seen intuitively, provided that the direction of the first proposal is chosen symmetrically and the sequence of jump lengths has the same distribution when reversed. Thus although the proposal density typically does not have a convenient closed form, it need not be evaluated in order to access the Metropolis acceptance probabilities. Another advantage is that the sampler is general-purpose, in the sense that no knowledge is required of the target density beyond the ability to evaluate it pointwise. In particular, it is not necessary to know the target's support a priori.

Beyond the context of random sampling, our work has applications to probabilistic methods for deterministic non-convex optimisation such as multistart Jain and Agogino (1993); Martí (2003) and basin-hopping Leary (2000); Wales and Doye (1997). These methods combine deterministic local search, such as given by a gradient method, with random perturbations or re-initialisations which may be performed using the skipping sampler to improve exploration. Section 5 provides numerical examples of these applications.

1.1 Related work

Many methods for accelerating the exploration of MCMC algorithms use prior knowledge of the target. For example, known mode locations may be used to design global moves for the sampler Andricioaei et al. (2001); Pompe et al. (2020);

Lan et al. (2014); Sminchisescu and Welling (2007); Sminchisescu et al. (2003); Tjelmeland and Hegstad (2001), or moves may be guided by the known derivatives of a differentiable target density Lan et al. (2014); Tak et al. (2018).

Some exceptions are methods that generate multiple proposals, such as Multipoint MCMC Qin and Liu (2001) and Multiple-try Metropolis Liu et al. (2000) which, like the skipping sampler, do not require additional information about the target. A fixed number of potentially correlated trial points are generated and one is selected at random, using a weight function which may be chosen to encourage exploration. Its random-grid implementation, in particular, has similarities with the skipping sampler. However, instead of fixing the number of draws, our proposal attempts to continue projecting further sequentially until it reaches A. Another advantage of our method is that it is Metropolis class, which simplifies both implementation and theoretical analysis.

During the review process our attention was brought to the very interesting sequential proposals of Park and Atchadé (2020), which also introduces a Metropolis-class sampler that modifies the proposal sequentially. In the wider class of algorithms introduced there, it is possible to recognise methods close in spirit to the skipping sampler. When skipping is applied to the hybrid slice sampler as in Sect. 5.2, for example, the resulting algorithm is a particular instance of the sequential proposal. While in Park and Atchadé (2020) the authors are motivated by the efficient implementation of Hamiltonian Monte Carlo, our own motivation is the efficient sampling of rare events. Together, these studies are suggestive of further potential to use sequences of proposals to accelerate MCMC methods in a range of situations, for instance within the framework introduced in Andrieu et al. (2020).

Like the hit-and-run sampler Smith (1984) and related algorithms (see Section 6.3 of Gilks and Roberts (1996)), the skipping sampler splits a Markovian transition into the random generation of a direction followed by a move in that direction. When the target conditioned on any line in the space is available in closed form, the hit-and-run algorithm is of course preferable, for the reasons provided in Rudolf and Ullrich (2018). Otherwise (which is more typical in applications), the skipping sampler offers a simple alternative and has the potential to increase exploration in the case of a non-convex support.

While also designed for targets with non-convex support, the *ghost sampler* introduced by the present authors in Moriarty et al. (2018) is not general-purpose since it uses knowledge of the geometry of the set A, assuming it is polyhedral.

The rest of the paper is structured as follows. We introduce the skipping sampler in Sect. 2 and state our main results in Sect. 3. Implementation and extensions are discussed in Sect. 4. Numerical applications to slice sampling and rare event
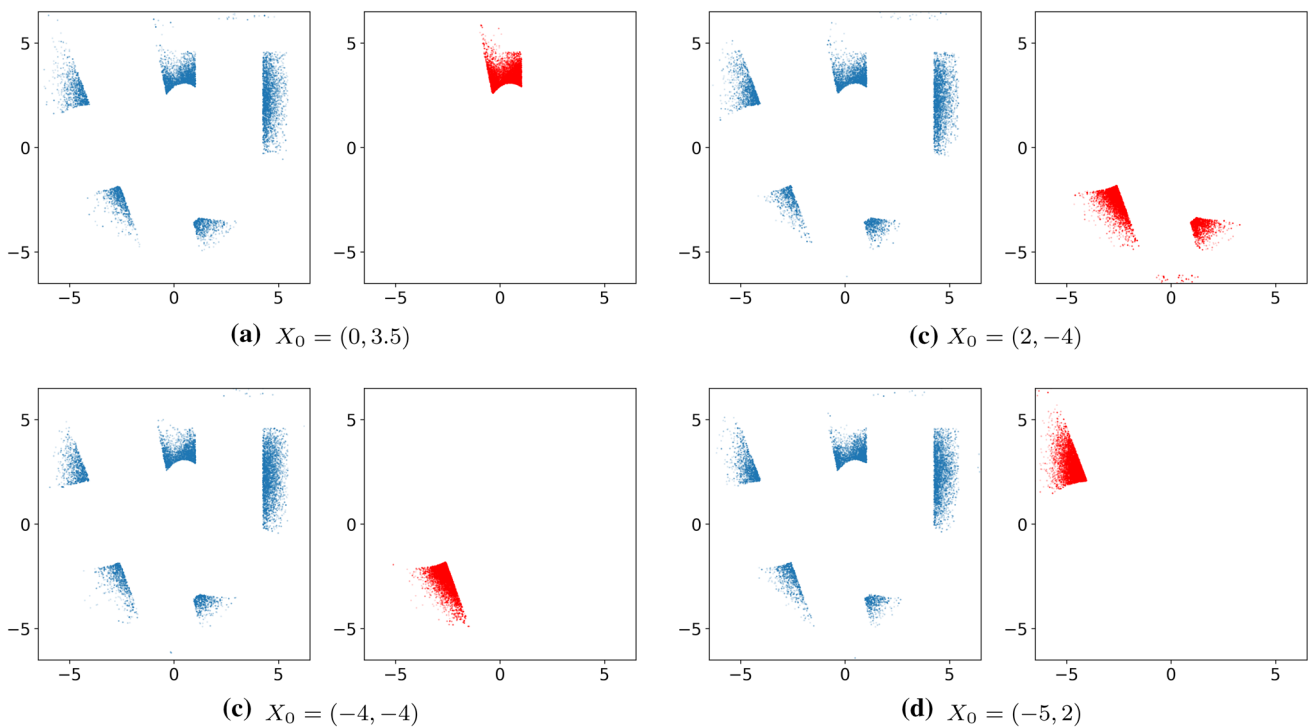
(a) $X_0 = (0, 3.5)$

(c) $X_0 = (2, -4)$

(c) $X_0 = (-4, -4)$

(d) $X_0 = (-5, 2)$

Fig. 1 Traces of the proposed skipping sampler (blue) and RWM algorithm (red) when the target has disconnected support. Both samplers are started at the same initial point X_0 and use the same underlying Gaussian proposal, whose standard deviation is tuned for a RWM acceptance ratio of 25%. The RWM typically localises around its initial state

sampling are given in Sect. 5, together with an application to global optimisation. Section 6 is devoted to the proof of the main results.

2 Skipping sampler

In this section we introduce the skipping sampler on $\mathbb{R}^d$, which is a modification of the RWM algorithm Metropolis et al. (1953). It is Metropolis-class although, unlike RWM, does not perform a random walk.

Assumption 2 Let $q : \mathbb{R}^d \to \mathbb{R}_+ \cup \{0\}$ be a symmetric $(q(x) = q(-x))$ continuous probability density function with $q(0) > 0$. We refer to q as the *underlying* proposal density.

Recall that given the state X_n of the chain, the RWM proposes a state Y_{n+1} sampled from the density $y \mapsto q(y - X_n)$ and accepts it as the next state X_{n+1} with probability

$$\alpha(X_n, Y_{n+1}) := \begin{cases} \min\left\{1, \frac{\pi(Y_{n+1})}{\pi(X_n)}\right\} & \text{if } \pi(X_n) \neq 0, \\ 1 & \text{otherwise,} \end{cases} \quad (1)$$

else it is rejected by setting $X_{n+1} = X_n$. Here π is the target density, although we do not take care to distinguish between π and the corresponding distribution as it will not

cause confusion. For convenience we use the common shorthand $\text{MH}(\pi, q)$ (after the more general Metropolis-Hastings algorithm, see Hastings (1970)) to refer to the Metropolis-class algorithm with target π and proposal q.

Algorithm 1 presents the skipping sampler, which aims to endow RWM with an improved ability to cross regions in which the target has zero density. Beginning with a RWM proposal Y_{n+1}, it continues jumping in a linear trajectory and accepts or rejects the first state of nonzero target density to be encountered. Thus any RWM proposal $Y_{n+1} \in A^c$, which would be rejected by $\text{MH}(\pi, q)$, is instead reused by adding jumps of random size in the direction $Y_{n+1} - X_n$ until either A is entered, or skipping is halted.

Algorithm 1 can be interpreted as follows. The *halting index* K is an independent random variable with distribution $\mathcal{K}$ on $\mathbb{Z}_{>0} \cup \{\infty\}$. If $K = 1$ then Y, the usual RWM proposal, is taken as the proposal. However if $K > 1$, the proposal is constructed using the *skipping chain* $\{Z_k\}_{k \geq 0}$ on $\mathbb{R}^d$ defined by $Z_0 := X$, with $X = X_n$ being the current state of the chain, and the update rule

$$Z_{k+1} := Z_k + \Phi R_{k+1}, \quad k \geq 0, \quad (3)$$

where $\|\cdot\|$ denotes the Euclidean norm, $\Phi = (Y - X)/\|Y - X\|$, $R_1 = \|Y - X\|$, and the distance increments $\{R_k\}_{k \geq 2}$

Algorithm 1: Skipping sampler (n-th iteration)

Input : The n-th sample $X_n \in \mathbb{R}^d$

1 Set $X := X_n$ and $Z_0 = X$;
2 Generate the initial proposal Y distributed according to the density $u \mapsto q(u - X)$;
3 Calculate the direction

$$\Phi = \frac{(Y - X)}{\|Y - X\|};$$

4 Generate an independent random halting index $K \sim \mathcal{K}$;
5 Set $k = 1$ and $Z_1 := Y$;
6 **while** $\underline{Z_k \in A^c}$ **and** $k < K$ **do**
7   Generate an independent distance increment R distributed as $\|Y - X\|$ given Φ;
8   Set $Z_{k+1} = Z_k + \Phi R$;
9   Increase k by one;
10 **end**
11 Set $Z := Z_k$;
12 Evaluate the acceptance probability:

$$\alpha(X, Z) = \begin{cases} \min\left(1, \frac{\pi(Z)}{\pi(X)}\right) & \text{if } \pi(X) \neq 0, \\ 1, & \text{otherwise}; \end{cases} \quad (2)$$

Generate a uniform random variable U on $(0, 1)$;
13 **if** $\underline{U \leq \alpha(X, Z)}$ **then**
14     $X_{n+1} = Z$;
15 **else**
16     $X_{n+1} = X$;
17 **end**
18 **return** $\underline{X_{n+1}}$.

are independent draws from the distribution of the radial part $\|Y - X\|$ conditional on the angular part Φ.

Let T_A be the first entry time of the skipping chain into A:

$$T_A := \min\{k \geq 1 : Z_k \in A\}, \quad (4)$$

with $\min \emptyset := \infty$. Writing $T_A \wedge K$ for the smaller of the two indices T_A and K, we also require:

Assumption 3 The support $A = \text{supp}(\pi)$ and distribution $\mathcal{K}$ are such that $\mathbb{E}[T_A \wedge K] < \infty$.

Relevant considerations for the choice of $\mathcal{K}$ and q are discussed in Sect. 4. Note that almost surely we have both $Y \neq x$ (since q is a density) and $T_A \wedge K < \infty$ (Assumption 3), so the skipping proposal $Z := Z_{T_A \wedge K}$ output by Algorithm 1 is well defined.

Proposition 1 *The following statements hold:*

(i) *Algorithm 1 is a symmetric Metropolis-class algorithm on the domain A. That is, there exists a transition density $q_{\mathcal{K}}$ (which depends on the halting index distribution $\mathcal{K}$) satisfying $q_{\mathcal{K}}(x, z) = q_{\mathcal{K}}(z, x)$ for all $x, z \in A$, such that Algorithm 1 is $MH(\pi, q_{\mathcal{K}})$.*

(ii) *The inequality $q_{\mathcal{K}}(x, z) \geq q(z-x)$ holds for every $x, z \in A$.*

Proof (i) We now make rigorous the intuitive argument which was provided earlier for the symmetry of the skipping proposal. Conditional on the direction Φ, the skipping chain (3) is one-dimensional. We therefore analyse this one-dimensional chain, before integrating over Φ to obtain the unconditional transition density.

Consider transitions of the skipping chain (3) between the states x and z in exactly $k \in \mathbb{Z}_{>0}$ steps. The intermediate states $z_1, \ldots, z_{k-1}$ satisfy $z_i \in A^c$ for $i = 1, \ldots, k - 1$. The (sub-Markovian) density $z \mapsto \xi_k(x, z)$ of these transitions is given by the Chapman-Kolmogorov equation and the density $\xi(r)$ of the distance increment R, which can in d-dimensional spherical coordinates be seen to be proportional to $q(r\Phi)r^{d-1}$. Since the distance increments are i.i.d. and have symmetric densities ($q(-r\Phi) = q(r\Phi)$), simple manipulations of the Chapman-Kolmogorov integral confirm that it is unchanged when the start and end point, the order of the jumps, and the direction of each jump are all reversed. This establishes that the density ξ_k is symmetric.

Next note that Assumption 3 implies the decomposition

$$\begin{aligned} \{Z = z\} &= \bigcup_{k=1}^{\infty} \{Z = z, \ T_A \leq k, \ K = k\} \\ &\cup \ \{Z = z, \ T_A < \infty, \ K = \infty\} \\ &\cup \bigcup_{k=1}^{\infty} \{Z = z, \ T_A > k, \ K = k\}. \end{aligned}$$

Hence, Z given x and Φ has a (sub-Markovian) density

$$\begin{aligned} \xi_{\mathcal{K}}(x, z) &= \sum_{k=1}^{\infty} \mathbb{P}[K = k] \sum_{j \leq k} \xi_j(x, z) 1_A(z) \\ &+ \ \mathbb{P}[K = \infty] \sum_{j=1}^{\infty} \xi_j(x, z) 1_A(z) \\ &+ \ \sum_{k=1}^{\infty} \mathbb{P}[K = k] \xi_k(x, z) 1_{A^c}(z). \end{aligned}$$

When $z \in A$ the above can be simplified to

$$\xi_{\mathcal{K}}(x, z) = \sum_{k=1}^{\infty} \xi_k(x, z) \, \mathbb{P}[K \geq k]. \quad (5)$$

Using d-dimensional spherical coordinates, the unconditional transition density is then the product of the density

of Φ with the transition density conditional on Φ:

$$q_{\mathcal{K}}(x, z) = \|z - x\|^{1-d} \xi_{\mathcal{K}}(x, z) \cdot \int_0^\infty$$
$$q\left(\frac{z - x}{\|z - x\|} r\right) r^{d-1} dr \,.$$

As the proposal q and the densities ξ_k (for all k) are symmetric, so is $\xi_{\mathcal{K}}$ and so is the skipping proposal $q_{\mathcal{K}}$, whenever $x, z \in A$.

Since any proposal $Z \in A^c$ is almost surely rejected if $x \in A$, Algorithm 1 is a well defined Metropolis-class algorithm on A, i.e. it is equivalent to $\mathrm{MH}(\pi, q_{\mathcal{K}})$ on the domain A.

(ii) As noted above, if $K = 1$ then Algorithm 1 reduces to $\mathrm{MH}(\pi, q)$. From (5) we therefore have

$$\xi_{\mathcal{K}}(x, z) = \sum_{k=1}^\infty \xi_k(x, z) \, \mathbb{P}[K \geq k] \geq \xi_1(x, z) \cdot 1$$

which again translates to the desired statement about proposal densities.

$\square$

3 Theoretical results

For completeness of the discussion below we provide the following definitions, further details of which may be found in Meyn and Tweedie (2009). A Markov chain $X_0, X_1 \dots$ is π-irreducible if for every $x \in \mathbb{R}^d$ and every $D \subset \mathbb{R}^d$ with $\pi(D) > 0$ we have

$$\mathbb{P}_x \left[\bigcup_{n \in \mathbb{Z}_{>0}} \{X_n \in D\} \right] > 0 \,.$$

Further, if $\mathbb{P}_x \left[\bigcup_{n \in \mathbb{Z}_{>0}} \{X_n \in D\} \right] = 1$ for every $x \in B$ and every $D \subset B$ with $\pi(D) > 0$ we say that $X_0, X_1, \dots$ is *Harris recurrent on B*. A set B is *absorbing* for a Markov chain with transition kernel P if $P(x, B) = 1$ holds for all $x \in B$. Note that an absorbing set B gives rise to a Markov chain evolving on B whose transition kernel is simply P restricted to B (see (Meyn and Tweedie 2009, Theorem 4.2.4)).

It is clear from (1) that if $x \in \mathrm{supp}(\pi)$ then

$$P(x, \mathrm{supp}(\pi)^c) = 0,$$

so that $\mathrm{supp}(\pi)$ is an absorbing set for the Metropolis algorithm with target π, and is a natural space of realisations of the chain. In what follows we therefore always consider the chain to evolve on the set A.

Regarding initialisation of the skipping sampler, note from (2) that if $X_0 \notin \mathrm{supp}(\pi)$ in Algorithm 1 then Z is automatically accepted. In this case the skipping sampler first enters $\mathrm{supp}(\pi)$ at a random step N and, for $0 \leq n \leq N - 2$, we have $X_{n+1} = Z_K$ – that is, the maximum allowed number of skips is performed at each stage. This implies that the skipping procedure is also capable of improving exploration in this initialisation stage. Theorem 1 assumes that $\pi(X_0) > 0$, or that initialisation has already been performed. We have

Theorem 1 *[SLLN] Suppose that $\mathrm{MH}(\pi, q)$ restricted to $\mathrm{supp}(\pi)$ is π-irreducible. Then $\mathrm{MH}(\pi, q_{\mathcal{K}})$ restricted to $A = \mathrm{supp}(\pi)$ is also π-irreducible and Harris recurrent. Moreover, the Strong Law of Large Numbers holds: if $\{X_i\}_{i \in \mathbb{Z}_{>0}}$ is the skipping sampler (generated by Algorithm 1) initiated at $X_0 = x \in A$, then for every π-integrable function f we have*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^n f(X_i) \overset{a.s.}{=} \int_{\mathbb{R}^d} f(x) \pi(x) dx \,.$$

The conditions of Theorem 1, which are mild, are discussed in Sect. 4. There are also cases where $\mathrm{MH}(\pi, q)$ is not irreducible but $\mathrm{MH}(\pi, q_{\mathcal{K}})$ is, for instance when the dimension $d = 1$, q is a random walk proposal with finite support, and A^c is an interval too wide to be crossed by a single random walk step, but which can be skipped across.

The statement of the second main result uses some additional notation (for further details see Roberts and Rosenthal (1997)). Consider the Hilbert space $L^2(\pi)$ of square-integrable functions with respect to π, equipped with the inner product (for $f, g \in L^2(\pi)$)

$$\langle f, g \rangle := \int_{\mathbb{R}^d} f(x) g(x) \pi(x) dx = \int_A f(x) g(x) \pi(x) dx.$$

Since all Metropolis-class chains are time reversible, the Markov kernel of $\mathrm{MH}(\pi, q)$ defines a bounded self-adjoint linear operator P on $L^2(\pi)$, defined for $f \in L^2(\pi)$ via

$$Pf(x) := \int_{\mathbb{R}^d} f(y) \alpha(x, y) q(y - x) dy$$
$$+ \left(1 - \int_{\mathbb{R}^d} \alpha(x, y) q(y - x) dy\right) f(x).$$

If P is irreducible then its operator norm is $\|P\| = 1$, with $f \equiv 1$ as the unique eigenfunction for the eigenvalue 1, and the spectral gap of P is defined to be $\lambda := 1 - \sup_{\{f \,:\, \|f\|=1, \, \pi(f)=0\}} \langle Pf, f \rangle$.

Theorem 2 *Under the conditions of Theorem 1, denoting respectively by P and $P_{\mathcal{K}}$ the Markov kernels of $\mathrm{MH}(\pi, q)$ and $\mathrm{MH}(\pi, q_{\mathcal{K}})$ restricted to $A = \mathrm{supp}(\pi)$, the following statements hold:*

(i) *For every* $f \in L^2(\pi)$ *we have* $\langle P_{\mathcal{K}} f, f \rangle \leq \langle P f, f \rangle$;

(ii) *If* $\mathrm{MH}(\pi, q)$ *has a non-zero spectral gap* λ, *then* $\mathrm{MH}(\pi, q_{\mathcal{K}})$ *also has a non-zero spectral gap* $\lambda_{\mathcal{K}}$ *that satisfies* $\lambda_{\mathcal{K}} \geq \lambda$;

(iii) *If the central limit theorem (CLT) holds for* $\mathrm{MH}(\pi, q)$ *and function* f *with asymptotic variance* $\sigma^2(f)$, *that is*

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=0}^{n} f(X_i) - \pi(f) \right) \longrightarrow N(0, \sigma^2(f))$$

in distribution, then the CLT also holds for $\mathrm{MH}(\pi, q_{\mathcal{K}})$ *and the same function* f, *with asymptotic variance* $\sigma_{\mathcal{K}}^2(f)$ *satisfying* $\sigma_{\mathcal{K}}^2(f) \leq \sigma^2(f)$.

The inequality at point (i) of Theorem 2 gives a useful way to compare performance and mixing of different Markov kernels. Indeed, one can consider the Peskun-Tierney partial ordering (see Peskun (1973) and Mira (2001); Mira and Leisen (2009); Tierney (1998)) on the family of bounded self-adjoint linear operators on $L^2(\pi)$ by setting $P_1 \geq P_2$ whenever $\langle P_1 f, f \rangle \leq \langle P_2 f, f \rangle$ holds for all $f \in L^2(\pi)$.

Intuitively, point (ii) of Theorem 2 means that the skipping sampler has the potential to mix faster than the classical random walk Metropolis, i.e., converge to stationarity in fewer steps. As explained in Sect. 2.1 of Rudolf and Ullrich (2018) the speed of convergence to stationarity can also be measured by other analytical quantities of the form $\inf_{f \in \mathcal{M}} \langle (I - P) f, f \rangle$ for some subset $\mathcal{M}$ of $L^2(\pi)$; it is straightforward to modify Theorem 2 accordingly. In the case of the spectral gap presented above we have $\mathcal{M} = \{f \in L^2(\pi) : \pi(f) = 0 \text{ and } \pi(f^2) = 1\}$. It follows from point (iii) that in stationarity, the samples produced by the skipping sampler are at least as good for estimating $\pi(f)$ as those generated by RWM.

These theoretical benefits are balanced by increased computational complexity. The exploration added by the skipping sampler relative to RWM carries a computational cost, and the tradeoff between cost and benefit depends on the target density. In particular, this tradeoff could become disadvantageous if evaluating the target density (and thus assessing the event $\{Z_k \in A\}$) in Algorithm 1 carries high cost. In the absence of global knowledge of the target, a pragmatic approach would be to run both methods and try to judge between their output. In Sect. 5.2, for example, we have compared the mean squared error of the coordinate projections against the increased number of evaluations of the target density. As noted in Sect. 4.1, the evaluations of the target density can also be vectorised with the aim of decreasing computation time.

Sufficient conditions for parts (ii) and (iii) of Theorem 2 have been studied in the literature. An aperiodic reversible Markov chain has non-zero spectral gap if and only if it is

geometrically ergodic (see Roberts and Rosenthal (1997)), a property which is explored in Jarner and Hansen (2000); Mengersen and Tweedie (1996); Roberts and Tweedie (1996) for random walk Metropolis algorithms. The CLT holds essentially for all $f \in L^2(\pi)$ under the assumption of geometric ergodicity (see (Roberts and Rosenthal 2004, Section 5)), but also holds more generally (see Jarner and Roberts (2002)).

4 Implementation and extensions

Implementing Algorithm 1 involves two choices, an underlying proposal density q and a halting index $\mathcal{K}$, which are discussed respectively in Sects. 4.1 and 4.2. An alternative to Algorithm 1 using a 'doubling trick' for greater computational efficiency is given in Sect. 4.3.

4.1 Choice of *q*

In addition to Assumptions 2–3, to ensure that the SLLN holds (Theorem 1) we require $\mathrm{MH}(\pi, q)$ to be π-irreducible. This holds, for example, when π is continuous and bounded and q is everywhere positive. More generally, $\mathrm{MH}(\pi, q)$ is also irreducible if the interior of A is (non-empty) connected and there exist $\delta, \epsilon > 0$ such that $q(x) > \epsilon > 0$ whenever $\|x\| < \delta$ (see (Tierney 1994, Section 2.3.2)).

Since skipping can be seen as a way of endowing RWM with an improved ability to cross regions of zero density, a minimal approach would be to tune q as if it were to be employed in the RWM algorithm, for example achieving an acceptance ratio around 25% when q is employed in RWM. However we have observed empirically that a lower acceptance ratio, for example 15%, may further stimulate skipping.

4.1.1 Computational aspects

For sampling of the i.i.d. radial increments $R_1, R_2, \ldots$, it is desirable to choose q such that samples may be drawn efficiently from

$$\|Y - X\| \text{ conditional on } \Phi = \frac{Y - X}{\|Y - X\|} = \varphi, \quad (6)$$

for all $\varphi \in \mathbb{S}^{d-1}$. Convenient cases include when q is radially symmetric so that conditioning is not required, or when $q \sim \mathcal{N}(0, \Sigma)$ for some $d \times d$ covariance matrix Σ, so that, given direction φ, each increment R_i follows a generalised gamma distribution with density

$$\frac{(\varphi^T \Sigma^{-1} \varphi)^{d/2}}{2^{d/2-1} \Gamma(\frac{d}{2})} r^{d-1} e^{-(\varphi^T \Sigma^{-1} \varphi) \frac{r^2}{2}}.$$

Alternatively one may specify q indirectly by choosing the unconditional distribution of Φ and the conditional distribution of $R := \|Y - X\|$ given Φ, then checking that the conditions of Theorem 1 are satisfied.

If sampling from the distribution (6) is computationally expensive, however, the sampler may be modified by setting all R_k equal to R, so that only a single sample is required to generate a proposal and the skipping chain keeps moving in the direction Φ with jumps of equal size. While this modification would not change the mean distance $\|Z_m - Z_0\| = \sum_{i=1}^{m} R_i$ covered by m steps of the skipping chain, it would increase its variance to $m^2 \mathrm{Var}(R)$.

4.1.2 Anisotropy

If A has a known anisotropy, the angular part of the underlying proposal may be chosen to favour certain directions in comparison to others, for example by tailoring the covariance matrix in a normal proposal $q \sim \mathcal{N}(0, \Sigma)$. This may be useful in high dimensional problems where otherwise, with high probability, the skipping chain may fail to re-enter A. It is not difficult to show that if the distance increment retains the properties used in the proof of Proposition 1, then the acceptance ratio (1) depends additionally on the ratio of the angular densities. Denoting by $q_\varphi(x, \phi)$ the density of direction ϕ at the location x, for $\Phi = \frac{Y_{n+1} - X_n}{\|Y_{n+1} - X_n\|}$ the acceptance probability then equals

$$\alpha(X_n, Y_{n+1}) \quad = \quad \min\left(1, \frac{\pi(Y_{n+1}) q_\varphi(Y_{n+1}, -\Phi)}{\pi(X_n) q_\varphi(X_n, \Phi)}\right).$$

Although beyond the scope of this paper, in the absence of geometric knowledge of A other information, for instance the history of the chain, may be used in an online fashion to make the angular part of the underlying proposal density dependent on the chain's current location.

4.2 Choice of $\mathcal{K}$

The simplest choice is a nonrandom halting index $K \equiv k_s \in \mathbb{Z}_{>1}$. Under this choice the k_s skips can be vectorised and stored in memory along with the corresponding states Z_i for $i = 1, \ldots k_s$, and the evaluations of whether $Z_i \in A$ for $i = 1, \ldots k_s$ can then be performed in parallel. This increases computational speed at the expense of a k_s-times higher memory requirement plus the coordination cost of parallelisation, and the balance between benefit and cost is not explored here. However if the additional computational costs are low, and if the costs of evaluating whether $Z_i \in A$ are bounded, then the skipping sampler may be run at speed approaching that of RWM.

There is of course interplay between the choices for $\mathcal{K}$ and q. For example, if an upper bound D is available for

the diameter of A^c then we may use $k_s = \frac{D}{\sup_\varphi \sigma_\varphi}$, where σ_φ denotes the standard deviation of the conditional jump density in the direction φ. In the anisotropic case of Section 4.1.2, *mutatis mutandis* the halting index may also be made direction-dependent using a parametric family of constants (or distributions) $\mathcal{K}_\varphi, \varphi \in \mathbb{S}^{d-1}$. To preserve symmetry it is then necessary that $\mathcal{K}_\varphi = \mathcal{K}_{-\varphi}$ for each $\varphi \in \mathbb{S}^{d-1}$. Similar tradeoffs between $\mathcal{K}$ and q may also be made when $\mathcal{K}$ is chosen to be random with finite mean.

If skipping cannot be efficiently parallelised as suggested above then, clearly, large realisations of $\mathcal{K}$ can result in high computational costs if A is not re-entered. In the extreme, bearing in mind Assumption 3, an unbounded distribution $\mathcal{K}$ should only be taken if A^c is known to be bounded. If $\mathcal{K}$ cannot be chosen based on a known diameter D as above, then the absolute length of skipping trajectories may alternatively be controlled probabilistically using a large deviations estimate, as follows. If the conditional jump distribution is R then the probability that a distance mr can be traversed in m skips is approximately (see for example Dembo and Zeitouni (2010)):

$$\mathbb{P}\left(\sum_{k=1}^{m} R_i \geq mr\right) \approx \exp(-m I(r)),$$

where $I(r) = \sup_{\theta>0}[\theta r - \lambda(\theta)]$ is the Legendre-Fenchel transform of R, provided that R has finite logarithmic moment generating function, i.e. $\lambda(\theta) = \ln \mathbb{E}[\exp(\theta R)] < \infty$ for all $\theta \in \mathbb{R}$.

Based on the above, if $\mathcal{K}$ is random and mass is to be placed on large values of $\mathcal{K}$ then this could lead to large computational costs. In this case the doubling trick of Section 4.3 may be applied.

4.3 The doubling trick

For clarity of exposition we first assume that A^c is convex. From (3), the state Z_k of the skipping chain is the partial sum $x + \Phi \sum_{i=1}^{k} R_i$, where the R_i are i.i.d. and $R_1 = \|Y - x\|$. Recalling (4), define

$$T_A := \min\{k \geq 1 : Z_k \in A\}, \tag{7}$$
$$S_A := \min\{k \geq 1 : Z_{2^k-1} \in A\}. \tag{8}$$

The convexity of A^c induces an ordering on the skipping chain, in the sense that

$$Z_k \in A^c, \quad \text{if } k < T_A, \tag{9}$$
$$Z_k \in A, \quad \text{if } k \geq T_A. \tag{10}$$

If $T_A < K$ then Algorithm 1 evaluates T_A by sampling the partial sums $\{Z_k\}_{k\geq1}$ sequentially. The following alternative

implementation evaluates T_A significantly faster, in order $\log_2 T_A$ steps. It requires that for any k, the sum $\sum_{i=1}^{k} R_i$ may be sampled directly, both unconditionally and given the value of $\sum_{i=1}^{2k} R_i$, at a comparable cost to sampling R_1. This is possible, for example, when the R_i are exponentially distributed.

The idea is to search forward through the exponential subsequence $Z_1, Z_3, Z_7 \ldots Z_{2^k-1}, \ldots$ until $k = \tilde{k} = S_A$ (so that $Z_{2^{\tilde{k}}-1} \in A$), and then to perform a logarithmic search Vijay-alakshmi Pai (2008) of the sequence $Z_{2^{\tilde{k}-1}-1}, \ldots, Z_{2^{\tilde{k}}-1}$ to identify T_A. That is, sample Z_m for $m = 2^{\tilde{k}-1} - 1 + 2^{\tilde{k}-2}$ and then, depending on whether or not it lies in A, reduce the search to either the sequence $Z_{2^{\tilde{k}-1}-1}, \ldots, Z_{2^{\tilde{k}-1}-1+2^{\tilde{k}-2}}$ or the sequence $Z_{2^{\tilde{k}-1}-1+2^{\tilde{k}-2}}, \ldots, Z_{2^{\tilde{k}}-1}$, repeating until T_A is found.

For generalisations of this trick, note first that the doubling trick can be used only to accelerate skipping over a convex subset $B \subset A^c$, so that we only add a single distance increment at a time while the skipping chain is in $A^c \setminus B$, and use the doubling trick while in B. The idea may then be applied to a maximal convex subset of A^c, provided that such a subset is known. Then note that if $B_1, \ldots, B_{n_B}$ are all convex subsets of A^c, the doubling trick may be used to traverse each convex subset B_i in turn, if needed. Thus the idea may be applied to an inner approximation of A^c by a union of balls, for example.

5 Numerical examples

In order to motivate some applications, Section 5.1 begins with a general discussion of targets for which the skipping sampler offers an advantage over RWM. The numerical example of Section 5.3, in the context of rare event sampling, illustrates an improvement in exploration achieved by our method. Then, in an application to optimisation, Section 5.4 provides quantitative examples of performance improvements obtained when the skipping sampler is used as a subroutine in probabilistic methods for non-convex optimisation. The Python code used to create all these numerical examples and figures is available at Zocca and Vogrinc (2021).

5.1 General considerations

Note firstly that if the initial proposal Y lies in A^c then it would be rejected by the RWM algorithm. Instead, in Algorithm 1 it is reused. Thus skipping offers an advantage over RWM if the initial proposal Y regularly lies in A^c. Secondly, when $Y \in A^c$ the skipping proposal Z of Algorithm 1 needs regularly to be accepted (which in turn necessitates $Z \in A$). By construction (since Z lies beyond Y on the straight line

between the current state $X_n \in A$ of the chain and $Y \in A^c$), this requires the support A of the target to be non-convex.

The dimension d also plays a key role. Considering an example where the support A is the union of two disjoint balls in $\mathbb{R}^d$, by increasing d we reduce the probability that $Z \in A$. Hence, the benefit of skipping is greatest in low dimensions and then gradually decreases. Nevertheless, in Sect. 5.3 we show that in special cases the sampler can be beneficial even in high dimensions.

We also note the following tradeoff. Due to the increased exploration offered by the skipping sampler, the density encountered upon landing at $Z \in A$ after crossing A^c may be significantly different from that at the current state X_n of the Markov chain. In particular, if the target density does not vary slowly then the acceptance ratio $\alpha(X, Z)$ may be so low that such skips are not regularly accepted. Although this tradeoff is problem dependent, it does not apply in the rare event example of Section 5.3.

5.2 Hybrid slice sampler

The slice sampler may be used to sample from a density ρ on $\mathbb{R}^d$ as follows. Given the current sample $X_n \in K$, the following two steps generate the next sample X_{n+1}:

(i) pick t uniformly at random from the interval $[0, \rho(X_n)]$,
(ii) sample uniformly from the 'slice' or superlevel set

$$A(t) := \{x \in K : \rho(x) \geq t\}.$$

We refer the reader to Łatuszyński and Rudolf (2014); Neal (2003) and references therein for more information on the slice sampler and its convergence properties.

Step (ii) is typically infeasible in multidimensional settings. Instead, in the *Hybrid Slice Sampler* (HSS) a Markov chain is used to approximately sample the uniform distribution on the slice. The following example illustrates the potential advantage of using the skipping sampler rather than RWM to generate this chain, since the slice may not be convex.

For ρ we take a uniform mixture of $m = 7$ standard normal densities in $d = 5$ dimensions, whose means are drawn uniformly at random from a box $B = [-12, 12]^5$. The underlying RWM proposal is a spherically symmetric Gaussian, with variance tuned to achieve an acceptance ratio of 23.5% in RWM. Independent trajectories (started in stationarity) of $n = 2 \cdot 10^5$ steps were generated for the HSS algorithm with respectively the RWM and the skipping sampler used to sample from the superlevel sets. The halting index is taken to be $\mathbb{P}[K_\varphi = 15] = 1$ for all $\varphi \in \mathbb{S}^{d-1}$.

As can be seen from Figure 2, the RWM implementation remains in the mode in which it was initiated. In contrast, the skipping sampler version transitions regularly between
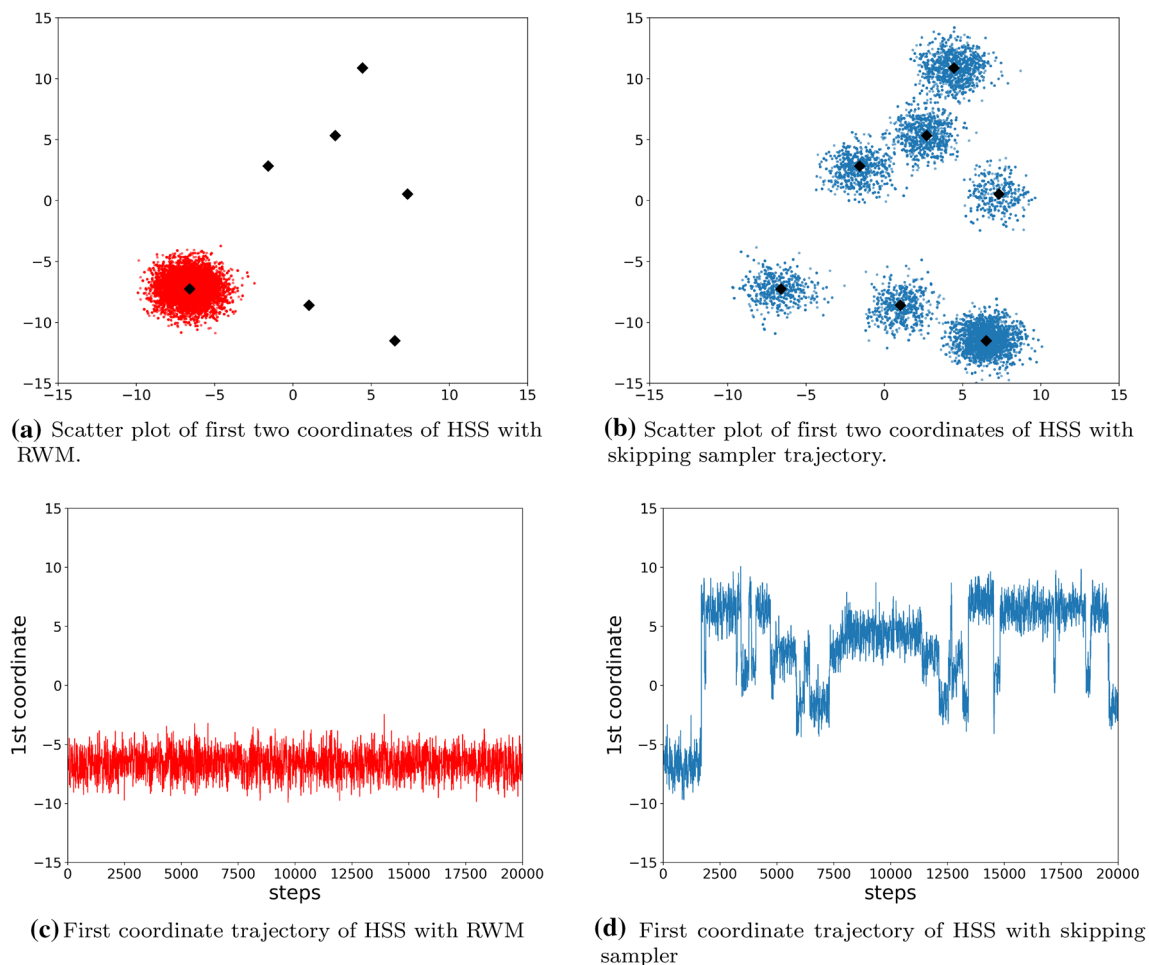
(a) Scatter plot of first two coordinates of HSS with RWM.

(b) Scatter plot of first two coordinates of HSS with skipping sampler trajectory.

(c) First coordinate trajectory of HSS with RWM

(d) First coordinate trajectory of HSS with skipping sampler

Fig. 2 Comparison of RWM and skipping sampler as subroutines in HSS when started at the same point. Black diamonds in **a** and **b** represent true mode centers

the seven modes. The experiment was run $m = 100$ times (on the same Gaussian mixture), during which skipping transitions happened on average 16 times per run. While the number of evaluations of the target density increased 11.61 fold on average, skipping greatly reduced the mean squared error (MSE) for the estimators of the coordinate-wise means. The MSE for RWM and reduction estimates (MSE for RWM divided by MSE for skipping) for each of the five coordinates are reported in Table 1. Hence, in this example the skipping sampler is roughly 12 times more expensive to compute, but produces samples with much greater effective sample size.

5.3 Rare event sampling

The aim in this example is to sample rare points under a complex density ρ on $\mathbb{R}^d$, by sampling from its intrinsic tail or sublevel set $A = \{x \in \mathbb{R}^d \mid \rho(x) \leq a\}$ for some $a > 0$. As an illustration let ρ be a mixture of $m = 20$ Gaussian distributions, with randomly drawn means, covariances and mixture coefficients.

Table 1 Mean squared errors for HSS with RWM or skipping sampler and its ratio

coordinate	1	2	3	4	5
MSE RWM	22.85	58.45	48.09	35.65	44.20
MSE Skipping	3.90	1.24	0.003	2.84	0.750
reduction	5.86	47.23	18504	12.54	58.91

We use the tails given by the levels $a = e^{-15}$ and $a = e^{-350}$ respectively for dimensions $d = 2$ and $d = 50$. In the case $d = 2$, a visual illustration of Theorem 2 is provided by plotting comparisons of the exploration achieved in 10^5 steps of RWM and the skipping sampler respectively. Since the superlevel sets of a finite Gaussian mixture are bounded, in this example we may take the halting index $K = \infty$.

Figure 3 a-b illustrates that, because of the density's exponential decay, samples from its tail are concentrated around the boundary ∂A of A. Figure 3 c-d compares the trajectories of the first coordinate of the chain, showing that while RWM

diffuses around ∂A, the skipping sampler regularly passes through A^c. Indeed, roughly 20% of the chain's increments were such 'skips' through A^c, almost half of the accepted proposals. The fact that proposals are often re-used rather than rejected is well illustrated by the acceptance rates, which are 23.7% and 43.3% for RWM and the skipping sampler respectively. Further, ∂A is disconnected. While the 'inner' component is not visited by RWM in this sample, the skipping sampler regularly passes through A^c to visit both components, thus exploring ∂A more quickly. Despite 3.45 times more target evaluations required for the skipping sampler, the benefits in this example are clearly worthwhile.

Figure 4 shows the evolution of the chain's first coordinate in the case $d = 50$. While the boundary of A cannot be easily visualised here, the faster mixing of the skipping sampler is again apparent. The successful re-use of proposals by skipping across A^c again constituted approximately 18% of the chain's steps, suggesting superdiffusive exploration. The respective acceptance rates were 22.2% for RWM and 48.1% for the skipping sampler. The benefits of skipping are again seen to be worth the computational cost, since this time the skipping sampler required only 1.44 times more target evaluations than RWM.

5.4 Applications to optimisation

The challenging problem of finding the global minimum of a non-convex function has attracted much attention and several probabilistic methods and heuristics have been developed, including simulated annealing Kirkpatrick et al. (1983), multistart Jain and Agogino (1993); Martí (2003), basin-hopping Leary (2000); Wales and Doye (1997), and random search Schumer and Steiglitz (1968). In this section we illustrate how the skipping sampler can be used in difficult low-dimensional examples to either bias the choice of initial points of such methods, or as a subroutine, in order to improve exploration. Below we consider an optimisation problem in $\mathbb{R}^d$ of the form

$$\min \quad f(\boldsymbol{x}) \quad \text{s.t.} \quad \boldsymbol{x} \in D := \prod_{i=1}^{d} [l_i, u_i], \tag{11}$$

and consider as the target density the Boltzmann distribution with temperature $T \geq 0$ and energy function f, conditioned on the region D, that is

$$\pi(\boldsymbol{x}) \propto \exp\left(-f(\boldsymbol{x})/T\right) \mathbf{1}_{\{\boldsymbol{x} \in D\}}. \tag{12}$$

5.4.1 Monotonic skipping sampler

While outside the scope of our theoretical analysis, a variation on Algorithm 1 is one in which the support A is not constant. In particular, defining the level sets $S(X_n) = \{\boldsymbol{x} \in$

$\mathbb{R}^d : f(\boldsymbol{x}) \leq f(X_n)\}$, a *monotonic skipping sampler (MSS)* may be defined in which the support at the n-th step of Algorithm 1 is $A_n := S(X_n) \cap D$ (setting $A_0 := D$), and the target density $\pi = \pi_n$ is uniform on A_n. That is, only downward moves (Markov chain transitions with $f(X_{n+1}) \leq f(X_n)$) are accepted. By construction we have $X_n \in A_n$ for each $n \in \mathbb{N}$. Also, since the random subsets $\{A_n\}_{n=1\ldots,m}$ are themselves decreasing with $A_{n+1} \subseteq A_n$ for every n, they contain progressively fewer non-global minima in addition to the global minima of the function f. In common with the skipping sampler where the support A is fixed, the n-th step of the MSS requires no information about the sublevel set $S(X_n)$, just the ability to check whether the proposal Z lies in A_n.

To illustrate a trajectory of the MSS, take f to be the so-called *eggholder function* in dimension $d = 2$, i.e.

$$f_{\text{eggholder}}(\boldsymbol{x}) := -x_1 \sin\left(\sqrt{|x_1 - x_2 - 47|}\right)$$
$$- (x_2 + 47) \sin\left(\sqrt{\left|\frac{x_1}{2} + x_2 + 47\right|}\right),$$

an optimisation test function often used in the literature Jamil and Yang (2013), with $D = [-512, 512]^2$. Figure 5 shows some snapshots from a trajectory of the MSS, also indicating the progressively shrinking sublevel sets $A_n = S(X_n) \cap D$. In this subsample the state of the chain (starred marker) is seen to jump four times between different connected components of the sublevel sets (in the subfigures for $n = 67$, 84, and 108), which happens by means of the skipping mechanism.

In Sections 5.4.2 and 5.4.3 we provide numerical examples of performance improvements achieved when the MSS is used as a subroutine in the multistart and basin-hopping optimisation procedures respectively.

5.4.2 Augmented multistart method

Given a nonconvex optimisation problem of the form (11) with possibly several local minima, a classical strategy to find its global minimum is to restart the local optimisation method of choice at several different points. The multistart method produces the desired number N of initial points by sampling them uniformly at random in $\prod_{i=1}^{d} [l_i, u_i]$.

Note that in the above setup, f may be set equal to positive infinity outside an arbitrary constraint set. If the set $f^{-1}(\mathbb{R})$ of feasible points has a low volume compared to D then many of the randomly sampled points may lie outside it, making this multistart initialisation procedure inefficient. In this case, recalling the remark on initialisation of Algorithm 1 from Section 2, the MSS is capable of accelerating the search for feasible starting points $\boldsymbol{x} \in f^{-1}(\mathbb{R})$.

Equally, sampling starting points uniformly at random may not be helpful if the basin of attraction of the global minimum has low volume. We can mitigate both of these issues
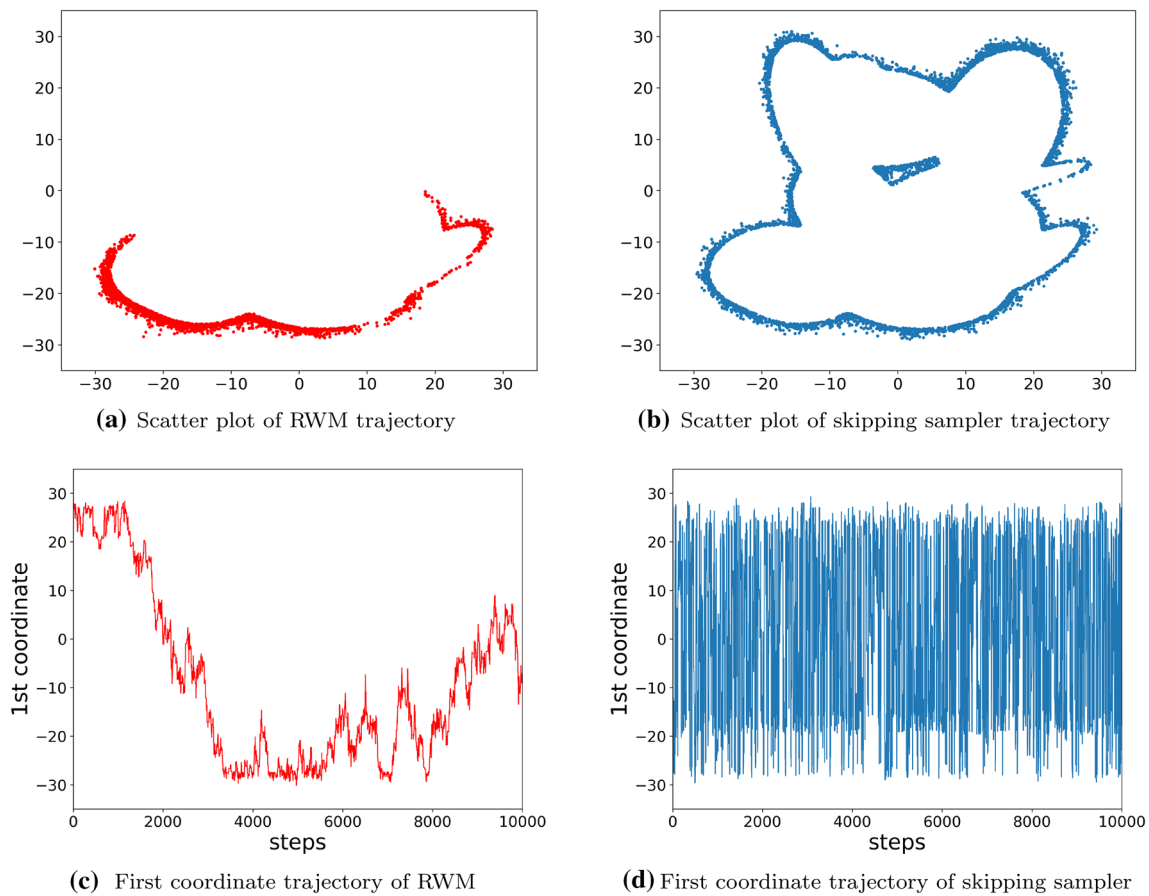
(a) Scatter plot of RWM trajectory

(b) Scatter plot of skipping sampler trajectory

(c) First coordinate trajectory of RWM

(d) First coordinate trajectory of skipping sampler

Fig. 3 Comparison of RWM and skipping sampler in dimension 2

by "improving" each of the points proposed by the multistart method as follows. Assume N initial points $X_0^{(1)}, \ldots, X_0^{(N)}$ have been sampled uniformly at random in $\prod_{i=1}^d [l_i, u_i]$, *which need not be feasible*. For each $i = 1, \ldots, N$, a Markov chain of length m may be generated using the MSS started at $X_0^{(i)}$, returning $X_m^{(i)}$. Algorithm 2 summarises in pseudo-code this MSS-augmented multistart method. By monotonicity, the augmented multistart procedure results in a greater proportion of feasible points, while each initially feasible point is improved.

Algorithm 2: MSS-augmented multistart method

Input : The number N of initial points and the desired length m of MSS trajectories

1　Generate N points uniformly at random $X_0^{(1)}, \ldots, X_0^{(N)}$ in $\prod_{i=1}^d [l_i, u_i]$;
2　**for** $i = 1$ **to** N **do**
3　　Starting at $X_0^{(i)}$ generate a trajectory of length m using the MSS
4　**end**
5　**return** the endpoints of the MSS trajectories $X_m^{(1)}, \ldots, X_m^{(N)}$.

To illustrate the potential of the MSS-augmented multistart method, we present an example again using the eggholder function. We first consider the unconstrained optimisation problem

$$\min \quad f_{\text{eggholder}}(\boldsymbol{x}) \quad \text{s.t.} \ \boldsymbol{x} \in [-512, 512]^2, \qquad (13)$$

which has the optimal solution $\boldsymbol{x}^* = (512, 404.2319)$, attaining the value $f_{\text{eggholder}}(\boldsymbol{x}^*) = -959.6407$. Averaging over $N = 1000$ runs, in Table 2 we summarise the "goodness" of the N starting points given by the following three methods:

(i)　multistart method, i.e., initial points uniformly distributed on $[-512, 512]^2$;

(ii)　the initial points obtained in (i) are evolved for $m = 100$ steps with a RWM with Gaussian proposals with covariance matrix $2 \cdot I$ and the Boltzmann distribution π in (12) with $T = 1.0$ as target distribution;

(iii)　the initial points obtained in (i) are evolved for $m = 100$ steps with the MSS using a deterministic halting index $\mathcal{K} = 200$ and the same Gaussian proposal density as in (ii).
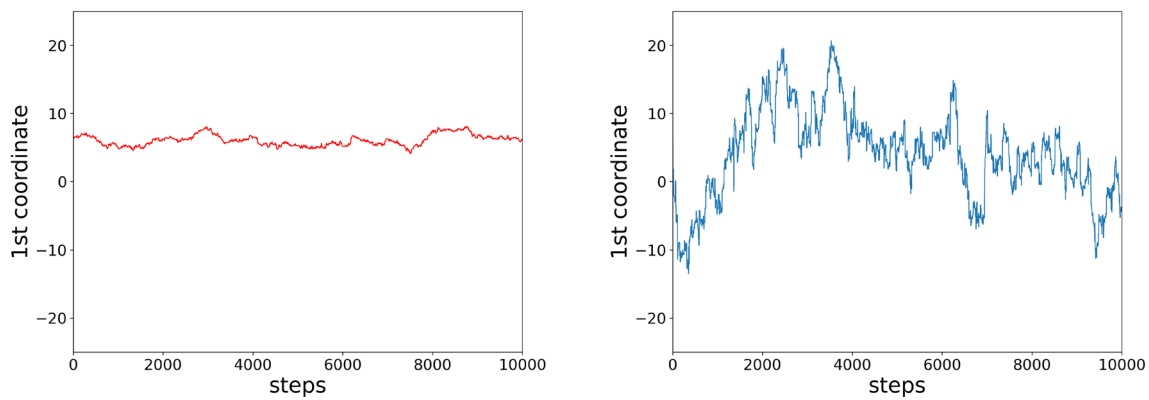
Fig. 4 First coordinate trajectory comparison for RWM (left) and skipping sampler (right) in dimension 50

Table 2 Comparison of the quality of the starting points returned by the three variants of the multistart method when solving the optimisation problem (13). The results are averaged over $N = 1000$ samples. The RWM and MSS variants both use trajectories of length $m = 100$ and the Gaussian proposal density $\mathcal{N}(\mathbf{0}, 2 \cdot I)$. The halting index for MSS was taken to be deterministic and equal to $\mathcal{K} = 200$. The acceptance probabilities for the RWM are evaluated w.r.t. the target distribution (12) with $T = 1.0$. For three of the metrics in the table, we report the median value and, in parenthesis, the 2.5 and 97.5 percentiles

Metric	Multistart method	Multistart method augmented with m RWM steps	Multistart method augmented with m MSS steps
Fraction of proposed points in the basin of attraction of the global minimum x^*	0.004	0.008	0.657
Euclidean distance from the global minimum x^*	741.85 (87.94, 1218.84)	741.85 (87.94, 1218.84)	0.0 (0.0, 977.869)
Gap between the optimum $f(x^*)$ and the objective value of the proposed point	465.68 (24.30, 848.49)	465.68 (24.30, 848.49)	0.0 (0.0, 70.69)
Number of function evaluations per run	36 (21, 84)	398 (325, 1909)	61527 (7805, 133470)
Running time (sec/run)	0.0021	0.0441	0.4577

The MSS augumented multistart method effectively biases the initial points towards the global minimum x^*, bringing 65.7% of them in the correct basin of attraction, although at the expense of more function evaluations than the other two methods.

5.4.3 Skipping sampler as basin-hopping subroutine

Besides improving the multistart method, the skipping sampler can also be used to improve stochastic techniques for non-convex optimisation, in particular the so-called *basin-hopping method*. In this subsection we explore this novel idea, although the implementation details and a systematic comparison with other global optimisation routines are left for future work.

Basin-hopping is a global optimisation technique proposed in Wales and Doye (1997), which at each stage combines a random perturbation of the current point, local optimisation, and an acceptance/rejection step. The random perturbation consists of i.i.d. uniform simultaneous perturbations in each of the coordinates, that is, a random walk

step. The stopping criterion for this iterative procedure is often a maximum number of function evaluations, or when no improvement is observed for a certain number of consecutive iterations.

The random walk step may be replaced by a step from the MSS. That is, at step n, given the current point X_{n-1} we first sample a new point Y_n from the sublevel set $S(X_{n-1}) \cap D$ using MSS and then perform a local optimisation procedure starting from Y_n to obtain a new point X_n. This idea is summarised in the pseudo-code presented in Algorithm 3.

The MSS variant of the basin-hopping method is related to the monotonic sequence basin-hopping (MSBH) proposed in Leary (2000), which also accepts only new points in $S(Y_{n-1}) \cap D$. However MSBH uses only local uniform perturbations and thus faces the same exploration challenges as RWM when $S(Y_{n-1}) \cap D$ is disconnected.

In Table 3, we compare the performance of basin-hopping and of the MSBH with the proposed basin-hopping with skipping. The MSS subroutine leads to 54.4% of the initial (uniformly distributed) points converging to the basin of attraction of the global minimum x^*. This sharp improvement
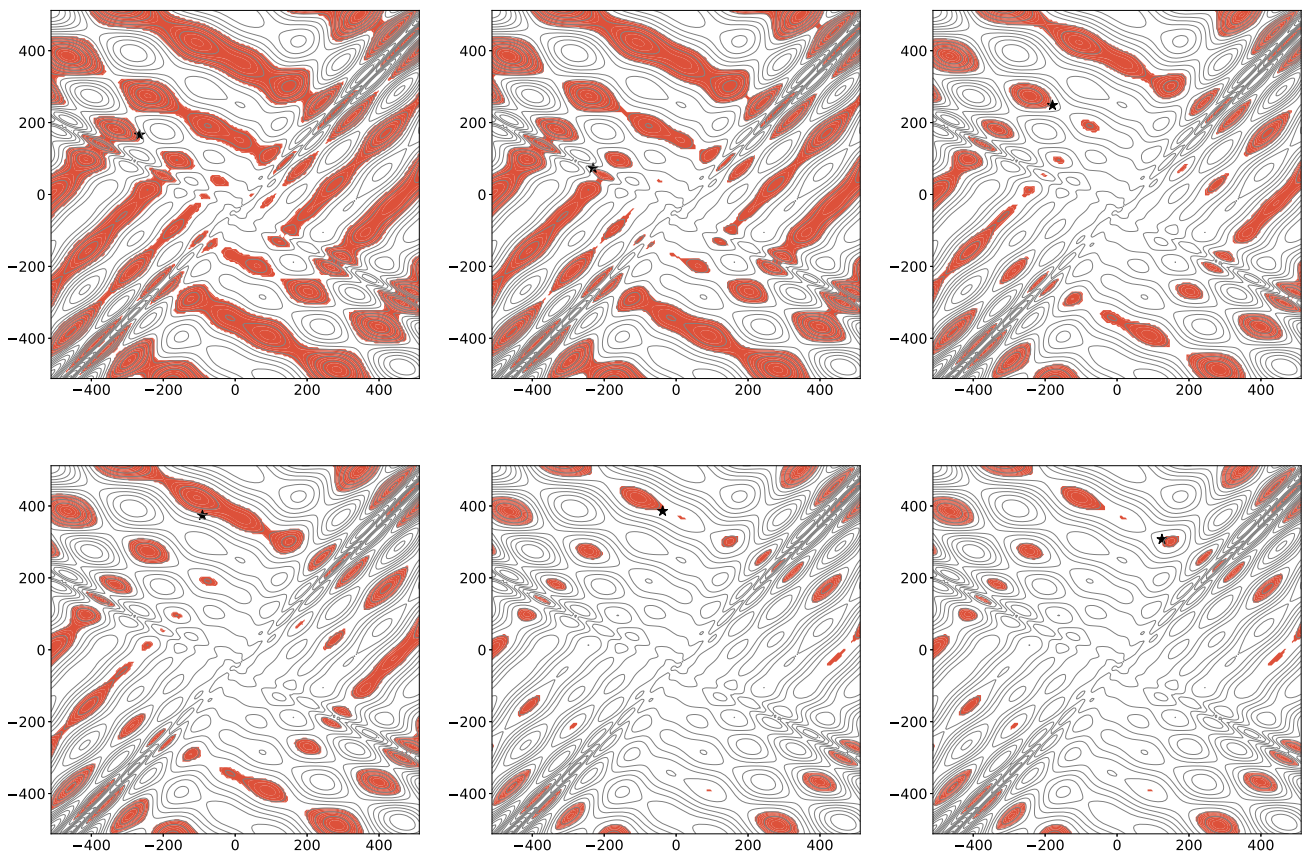
Fig. 5 Given a trajectory $(X_n)_{n=0,1,...}$ of the MSS started at $X_0 = (-200, 180)$, each subfigure displays the point X_n (starred marker) and corresponding sublevel set $S(X_n)$ (in red) for $n = 66, 67, 83$ (first row) and $n = 84, 107, 108$ (second row). There are a total of 54 skipping moves in this trajectory, which corresponds to 36.0% of the moves. The MSS uses the Gaussian distribution $\mathcal{N}(\mathbf{0}, 2 \cdot I)$ as proposal density and a deterministic halting index $\mathcal{K} = 150$

Table 3 Performance comparison of the three basin-hopping variants using different subroutines averaged over $N = 1000$ samples for trajectories of length $m = 100$. The underlying proposal used for the MSS subroutine is a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, I)$ and the uniform displacement of the other two methods is scaled to have the same standard deviation. The halting index for MSS was taken to be deterministic and equal to $\mathcal{K} = 200$. The acceptance probabilities for the basin-hopping methods are evaluated w.r.t. the target distribution (12) with $T = 1.0$. For three of the metrics in the table, we report the median value and, in parenthesis, the 2.5 and 97.5 percentiles

Metric	Basin-hopping (with uniform displacement as subroutine)	Monotonic sequence basin-hopping (with uniform displacement as subroutine)	Basin-hopping with skipping
Fraction of proposed points in the basin of attraction of the global minimum x^*	0.022	0.017	0.544
Euclidean distance from global minimum x^*	741.85 (87.94, 1248.15)	752.827 (87.94, 1248.15)	0.0 (0.0, 41.23)
Gap between the optimum $f(x^*)$ and the objective value of the proposed point	419.21 (24.30, 835.76)	402.03 (24.30, 835.85)	0.0 (0.0, 2.72)
Number of function evaluations	2004 (1776, 3432)	348 (324, 5085)	20370 (19957, 23702)
Running time (sec/run)	0.123	0.065	0.333

Algorithm 3: Basin-hopping with skipping

Input : An initial point $X_0 \in \prod_{i=1}^{d} [l_i, u_i]$, not necessarily feasible

1 Set $n = 1$;
2 **while** <u>basin-hopping stopping criterion is not satisfied</u> **do**
3   Perform a single step of the MSS started at X_{n-1}, obtaining a new point Y_n;
4   Perform a local optimisation step started at Y_n to obtain a new point X_n;
5   Increment the index n by one;
6 **end**
7 **return** <u>The last point X_n.</u>

---

with respect to basin-hopping (which has a corresponding success rate of only 2.2%) only requires ten times more function evaluations. In Goodridge et al. (2021) we present additional performance metrics for the basin-hopping method with skipping and more extensive numerical results over a large collection of test functions.

6 Proofs

6.1 Proof of Theorem 1

For each $x \in \text{supp}(\pi)$ let $\mathbb{P}_x$ be a probability measure carrying all random variables used in Algorithm 1, such that $X_0 = x$ almost surely under $\mathbb{P}_x$. Denote respectively by $\{Y_m\}_{m \geq 1}$ and $\{X_n\}_{n \geq 1}$ the proposals generated by the $\text{MH}(\pi, q)$ algorithm and the Markov chain returned by the algorithm. Writing $\mathcal{A}_n := \bigcap_{i=1}^{n}\{X_i = Y_i\}$ for the event that the first n proposals of $\text{MH}(\pi, q)$ are all accepted, we have

Lemma 1 *If the chain* $\text{MH}(\pi, q)$ *restricted to* $\text{supp}(\pi)$ *is* π-*irreducible then* $\mathbb{P}_x(\mathcal{A}_m) > 0$ *for all* $x \in supp(\pi)$ *and all* $m \geq 1$.

Proof Fixing $x \in \text{supp}(\pi)$ and supposing otherwise for a contradiction, let n be the smallest integer such that $\mathbb{P}_x(\mathcal{A}_n) = 0$. Clearly $n \geq 2$, since otherwise, $\mathbb{P}_x-$almost surely we have $X_k = X_0$ for all $k \geq 1$, contradicting the assumption of π-irreducibility. Therefore $\mathbb{P}_x(\mathcal{A}_{n-1}) > 0$ and we may write p for the density of X_{n-1} conditional on the event $\mathcal{A}_{n-1}$. Then by the Markov property we have

$$
\begin{aligned}
0 &= \mathbb{P}_x(\mathcal{A}_{n-1}) \, \mathbb{P}_x(\mathcal{A}_n | \mathcal{A}_{n-1}) \\
&= \mathbb{P}_x(\mathcal{A}_{n-1}) \int_{\text{supp}(p)} p(y) \, \mathbb{P}_y(\mathcal{A}_1) \, dy,
\end{aligned}
$$

so that $\mathbb{P}_y(\mathcal{A}_1) = 0$ for some $y \in \text{supp}(p)$. Arguing as above, this contradicts the assumption of π-irreducibility. $\qquad\square$

Denote the Markov kernels of the chains generated by $\text{MH}(\pi, q)$ and $\text{MH}(\pi, q_{\mathcal{K}})$ by P and $P_{\mathcal{K}}$ respectively. Also

let $\{X'_n\}_{n \geq 1}$ be the jump chain associated with X (that is, the subsequence of $\{X_n\}_{n \geq 1}$ given by excluding all X_m which satisfy $X_m = X_{m-1}$).

Lemma 2 *For all* $x \in A = \text{supp}(\pi)$, $n \in \mathbb{Z}_{>0}$ *and all* $B \subset A$ *the following inequality holds:*

$$
\begin{aligned}
P_{\mathcal{K}}^n(x, B) &\geq \mathbb{P}_x(\{X_n \in B\} \cap \mathcal{A}_n) \\
&= \mathbb{P}_x(X_n \in B \mid \mathcal{A}_n) \, \mathbb{P}_x(\mathcal{A}_n) \\
&= \mathbb{P}_x(X'_n \in B \mid \mathcal{A}_n) \, \mathbb{P}_x(\mathcal{A}_n). \quad (14)
\end{aligned}
$$

Proof Note first that the last equality in (14) follows by definition of the jump chain. We will prove the inequality in (14) by induction on n. Since $\text{supp}(\pi) = A$, Proposition 1 (ii) gives

$$
\begin{aligned}
P_{\mathcal{K}}(x, B) &\geq \int_B \alpha(x, z) q_{\mathcal{K}}(x, z) dz \\
&\geq \int_B \alpha(x, z) q(z - x) dz \\
&= \mathbb{P}_x(\{X_1 \in B\} \cap \mathcal{A}_1).
\end{aligned}
$$

Assume now the statement holds for some $n \in \mathbb{Z}_{>0}$ and let us prove it for $n+1$. We argue using the induction hypothesis and Proposition 1 (ii) again:

$$
\begin{aligned}
P_{\mathcal{K}}^{n+1}(x, B) &= \int_A P_{\mathcal{K}}^n(z, B) P_{\mathcal{K}}(x, dz) \\
&\geq \int_A P_{\mathcal{K}}^n(z, B) \alpha(x, z) q_{\mathcal{K}}(x, z) \, dz \\
&\geq \int_A P_{\mathcal{K}}^n(z, B) \alpha(x, z) q(z - x) \, dz \\
&\geq \int_A \mathbb{P}_z(\{X_n \in B\} \cap \mathcal{A}_n) \, \alpha(x, z) q(z - x) \, dz \\
&= \mathbb{P}_x(\{X_{n+1} \in B\} \cap \mathcal{A}_{n+1}).
\end{aligned}
$$

Proof of Theorem 1 Take $B \subseteq A = \text{supp}(\pi)$ such that $\pi(B) > 0$, $x \in A$ and let $\{X_n\}_{n \geq 1}$ be $\text{MH}(\pi, q)$ started at $X_0 = x$. Since $\text{MH}(\pi, q)$ is π-irreducible there exists an integer $n \in \mathbb{Z}_{>0}$ such that $\mathbb{P}_x(X_n \in B) > 0$. Let S_n be the number of rejections which occur in the generation of $\{X_m\}_{1 \leq m \leq n}$. Then

$$
0 < \mathbb{P}_x(X_n \in B) = \sum_{i=0}^{n} \mathbb{P}_x(X_n \in B, S_n = i).
$$

For some $j \in \{1, \dots, n\}$ we therefore have

$$
\mathbb{P}_x(X_n \in B, S_n = j) > 0.
$$

Consequently

$$
\mathbb{P}_x\left(X'_{n-j} \in B | \mathcal{A}_{n-j}\right) > 0,
$$

so that

$$P_{\mathcal{K}}^{n-j}(x, B) \geq \mathbb{P}_x\left(X'_{n-j} \in B \mid \mathcal{A}_{n-j}\right)\mathbb{P}_x\left(\mathcal{A}_{n-j}\right) > 0,$$

where we used the above together with Lemma 2 and Lemma 1. The skipping chain MH$(\pi, q_{\mathcal{K}})$ is therefore π-irreducible, and thus is Harris recurrent by (Tierney 1994, Corollary 2). Furthermore, (Meyn and Tweedie 2009, Theorem 10.0.1) yields that π is its unique invariant probability measure. Finally, the SLLN holds for all π-integrable functions by Harris recurrence and (Meyn and Tweedie 2009, Theorem 17.1.7). □

6.2 Proof of Theorem 2

To prove Theorem 2, we will make use of the following lemma, whose proof is omitted.

Lemma 3 *(Integration with respect to a symmetric joint density) Consider a symmetric density $\Delta : \mathbb{R}^d \times \mathbb{R}^d \to [0, +\infty)$ and a subset $B \subseteq \mathbb{R}^d$. For every $f \in L^2(\Delta)$ the following identity holds:*

$$\int_B \int_B \frac{f(x)^2 + f(y)^2}{2}\Delta(x, y)dydx$$
$$= \int_B f(x)^2\left(\int_B \Delta(x, y)dy\right)dx.$$

Proof of theorem 2 (i) For any $f \in L^2(\pi)$ the desired inequality $\langle P_{\mathcal{K}}f, f\rangle \leq \langle Pf, f\rangle$ can be written more explicitly as

$$\int_{\mathbb{R}^d} f(x)\Bigg(\left(\int_{\mathbb{R}^d} f(y)\alpha(x, y)(q_{\mathcal{K}}(x, y) - q(y - x))dy\right)$$
$$+ f(x)(r_{\mathcal{K}}(x) - r(x))\Bigg)\pi(x)dx \leq 0,$$

where we respectively denote by $r(x)$ and $r_{\mathcal{K}}(x)$ the rejection probabilities starting at point x of MH(ρ, q) and MH$(\rho, q_{\mathcal{K}})$, i.e., $r(x) := 1 - \int_{\mathbb{R}^d}\alpha(x, y)q(y - x)dy$ and analogously for $r_{\mathcal{K}}(x)$. The above inequality holds provided that we establish the following one:

$$\int_{\mathbb{R}^d} f(x)\left(\int_{\mathbb{R}^d} f(y)\alpha(x, y)\left(q_{\mathcal{K}}(x, y) - q(y - x)\right)dy\right)\pi(x)dx$$
$$\leq \int_{\mathbb{R}^d} f^2(x)(r(x) - r_{\mathcal{K}}(x))\pi(x)dx. \quad (15)$$

Then considering the LHS of (15) and Proposition 1 (ii) we have:

$$\int_{\mathbb{R}^d} f(x)\left(\int_{\mathbb{R}^d} f(y)\alpha(x, y)(q_{\mathcal{K}}(x, y) - q(y - x))dy\right)\pi(x)dx$$

$$= \int_A\int_A f(y)f(x)\alpha(x, y)\pi(x)(q_{\mathcal{K}}(x, y) - q(y - x))\,dydx$$
$$\leq \int_A\int_A \frac{f^2(y) + f^2(x)}{2}\alpha(x, y)\pi(x)(q_{\mathcal{K}}(x, y) - q(y - x))\,dydx$$
$$\stackrel{(\star)}{=} \int_A\int_A f(x)^2\alpha(x, y)\pi(x)(q_{\mathcal{K}}(x, y) - q(y - x))\,dydx$$
$$= \int_A f(x)^2\left(\int_A \alpha(x, y)(q_{\mathcal{K}}(x, y) - q(y - x))\,dy\right)\pi(x)dx$$
$$= \int_{\mathbb{R}^d} f^2(x)(r(x) - r_{\mathcal{K}}(x))\pi(x)dx.$$

In this derivation we used (in order) the fact that $\alpha(x, y) = 0$ for $y \in A^c$ by definition of α and π and the classical GM-QM inequality $2f(x)f(y) \leq f(x)^2 + f(y)^2$. Furthermore, equality ($\star$) holds thanks to Lemma 3 by taking $\Delta(x, y) = \alpha(x, y)(q_{\mathcal{K}}(x, y) - q(y - x))\pi(x)$ and $B = A$. The property that $\Delta(x, y) = \Delta(y, x)$ for every $x, y \in A$ readily follows by combining the following two identities that hold for every $x, y \in A$:

$$\alpha(x, y)\pi(x) = \min(\pi(x), \pi(y)) = \alpha(y, x)\pi(y), \quad \text{and}$$
$$q_{\mathcal{K}}(x, y) - q(y - x) = q_{\mathcal{K}}(y, x) - q(x - y).$$

The first identity is an immediate consequence of the definition (2) of α, while the second one follows from Assumption 2 and Proposition 1 (i).

(ii) By (i) we have $\langle(I - P_{\mathcal{K}})f, f\rangle \geq \langle(I - P)f, f\rangle$ for all $f \in L^2(\pi)$. The proof follows by $\lambda_{\mathcal{K}} = \inf_{f \in \mathcal{M}}\langle(I - P_{\mathcal{K}})f, f\rangle \geq \inf_{f \in \mathcal{M}}\langle(I - P)f, f\rangle = \lambda$ where $\mathcal{M} = \{f \in L^2(\pi) : \pi(f^2) = 1, \pi(f) = 0\}$.

(iii) This follows by (i) and (Mira and Leisen 2009, Theorem 6). □

References

Andricioaei, I., Straub, J., Voter, A.: Smart darting Monte Carlo. J. Chem. Phys. **114**(16), 6994–7000 (2001)

Andrieu, C., Lee, A., Livingstone, S.: A general perspective on the Metropolis-Hastings kernel. arXiv preprint arXiv:2012.14881 (2020)

Dembo, A., Zeitouni, O.: Large Deviations Techniques and Applications. Springer, Berlin Heidelberg (2010). https://doi.org/10.1007/978-3-642-03311-7

Gilks, W.R., Roberts, G.O.: Strategies for improving MCMC. Markov chain Monte Carlo in practice **6**, 89–114 (1996)

Goodridge, M., Moriarty, J., Vogrinc, J., Zocca, A.: Hopping between distant basins. arXiv preprint arXiv:2108.05229 (2021)

Hastings, W.: Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**(1), 97–109 (1970). https://doi.org/10.1093/biomet/57.1.97

Jain, P., Agogino, A.M.: Global optimization using the multistart method. J. Mech. Design **115**(4), 770–775 (1993). https://doi.org/10.1115/1.2919267

Jamil, M., Yang, X.S.: A literature survey of benchmark functions for global optimisation problems. Int. J. Math. Modell. Numerical Optimisation **4**(2), 150 (2013). https://doi.org/10.1504/ijmmno.2013.055204

Jarner, S., Hansen, E.: Geometric ergodicity of metropolis algorithms. Stochastic Process. their Appl. **85**(2), 341–361 (2000). https://doi.org/10.1016/S0304-4149(99)00082-4

Jarner, S., Roberts, G.: Polynomial convergence rates of Markov chains. Ann. Appl. Probab. **12**(1), 224–247 (2002)

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. Science **220**(4598), 671–680 (1983). https://doi.org/10.1126/science.220.4598.671

Lan, S., Streets, J., Shahbaba, B.: Wormhole Hamiltonian Monte Carlo. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence, pp. 1953–1959 (2014)

Łatuszyński, K., Rudolf, D.: Convergence of hybrid slice sampling via spectral gap (2014)

Leary, R.H.: Global Optimization on Funneling Landscapes. J. Glob. Optimiz. **18**(4), 367–383 (2000). https://doi.org/10.1023/a:1026500301312

Liu, J., Liang, F., Wong, W.: The multiple-try method and local optimization in Metropolis sampling. J. Am. Statistical Association **95**(449), 121–134 (2000)

Martí, R.: Multi-start methods. In: Handbook of Metaheuristics, pp. 355–368. Kluwer Academic Publishers (2003). https://doi.org/10.1007/0-306-48056-5_12

Mengersen, K., Tweedie, R.: Rates of convergence of the Hastings and Metropolis algorithms. Ann. Statistics **24**(1), 101–121 (1996). https://doi.org/10.1214/aos/1033066201

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E.: Equation of state calculations by fast computing machines. J. Chem. Phys. **21**(6), 1087–1092 (1953)

Meyn, S., Tweedie, R.: Markov chains and stochastic stability, 2nd edn. Cambridge University Press, Cambridge (2009). https://doi.org/10.1017/CBO9780511626630

Mira, A.: Ordering and improving the performance of Monte Carlo Markov chains. Statistical Science , 340–350 (2001)

Mira, A., Leisen, F.: Covariance ordering for discrete and continuous time Markov chains. Statistica Sinica **19**(2), 651–666 (2009)

Moriarty, J., Vogrinc, J., Zocca, A.: Frequency violations from random disturbances: an MCMC approach. In: IEEE Conference on Decision and Control (CDC), pp. 1598–1603 (2018). https://doi.org/10.1109/CDC.2018.8619304

Neal, R.: Slice sampling. Ann. Statistics **31**(3), 705–767 (2003). https://doi.org/10.1214/aos/1056562461

Park, J., Atchadé, Y.: Markov chain monte carlo algorithms with sequential proposals. Statistics Comput. **30**(5), 1325–1345 (2020). https://doi.org/10.1007/s11222-020-09948-4

Peskun, P.: Optimum Monte-carlo sampling using Markov chains. Biometrika **60**(3), 607–612 (1973)

Pompe, E., Holmes, C., Łatuszyński, K.: A framework for adaptive MCMC targeting multimodal distributions. The Annals of Statistics **48**(5),(2020). https://doi.org/10.1214/19-aos1916

Qin, Z., Liu, J.: Multipoint Metropolis method with application to hybrid Monte Carlo. J. Comput. Phys. **172**(2), 827–840 (2001)

Robert, C.P., Elvira, V., Tawn, N., Wu, C.: Accelerating MCMC algorithms. Wiley Interdisciplinary Reviews: Computational Statistics **10**(5), e1435 (2018)

Roberts, G., Rosenthal, J.: Geometric ergodicity and hybrid Markov chains. Electron. Commun. Probab. **2**, 13–25 (1997). https://doi.org/10.1214/ECP.v2-981

Roberts, G., Rosenthal, J.: General state space Markov chains and MCMC algorithms. Probab. Surv. **1**, 20–71 (2004). https://doi.org/10.1214/154957804100000024

Roberts, G., Tweedie, R.: Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. Biometrika **83**(1), 95–110 (1996). https://doi.org/10.1093/biomet/83.1.95

Rudolf, D., Ullrich, M.: Comparison of hit-and-run, slice sampler and random walk metropolis. J. Appl. Probab. **55**(4), 1186–1202 (2018)

Schumer, M., Steiglitz, K.: Adaptive step size random search. IEEE Trans. Automatic Control **13**(3), 270–276 (1968). https://doi.org/10.1109/tac.1968.1098903

Sminchisescu, C., Welling, M.: Generalized Darting Monte Carlo. In: M. Meila, X. Shen (eds.) Proceedings of the 11th International Conference on Artificial Intelligence and Statistics, vol. 2, pp. 516–523. PMLR (2007)

Sminchisescu, C., Welling, M., Hinton, G.: A mode-hopping MCMC sampler. Tech. rep., CSRG-478, University of Toronto (2003)

Smith, R.L.: Efficient monte carlo procedures for generating points uniformly distributed over bounded regions. Operations Res. **32**(6), 1296–1308 (1984)

Tak, H., Meng, X.L., van Dyk, D.: A Repelling-Attracting Metropolis Algorithm for Multimodality. Journal of Computational and Graphical Statistics (2018). https://doi.org/10.1080/10618600.2017.1415911

Tierney, L.: Markov chains for exploring posterior distributions. The Annals of Statistics **22**(4), 1701–1762 (1994). https://doi.org/10.1214/aos/1176325750. With discussion and a rejoinder by the author

Tierney, L.: A note on Metropolis-Hastings kernels for general state spaces. Ann. Appl. Probab. **8**(1), 1–9 (1998). https://doi.org/10.1214/aoap/1027961031

Tjelmeland, H., Hegstad, B.: Mode jumping proposals in MCMC. Scandinavian J. Statistics **28**(1), 205–223 (2001)

Vijayalakshmi Pai, G.: Data Structures and Algorithms: Concepts. Techniques and Applications, McGraw-Hill Education (2008)

Wales, D.J., Doye, J.P.K.: Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. J. Phys. Chem. A **101**(28), 5111–5116 (1997). https://doi.org/10.1021/jp970984n

Zocca, A., Vogrinc, J.: Skipping sampler code. https://github.com/alessandrozocca/skippingsampler (2021)