Non-stationary phase of the MALA algorithm with annealed proposals

Mylène BÉDARD [*]

May 2024

Abstract

The Metropolis-adjusted Langevin algorithm (MALA) is an informed MCMC method that is used to sample from a target distribution of interest. Its proposal distribution makes use of the gradient of the target's log-density in order to generate suitable candidates for the chain. This sampler is quite efficient in the stationary phase, but displays a notoriously erratic behaviour out of stationarity.

The Metropolis-adjusted Langevin algorithm with annealed proposals (aMALA) is a generalization of the usual MALA that features two tuning parameters: the usual step size δ and a parameter γ that may be adjusted to accommodate the dimension of the target distribution (with $\gamma = 1$ corresponding to MALA). It has been established in Boisvert-Beaudry and Bédard (2022) that aMALA with $1 < \gamma \leq 2$ usually outperforms MALA, even in high-dimensional contexts where the latter should become optimal.

The results of this paper demonstrate that the computational cost of aMALA is $\mathcal{O}(N^{1/3})$ in its non-stationary regime and that it may be as small as $\mathcal{O}(N^{1/5})$ in stationarity. This is in contrast to MALA, whose cost is $\mathcal{O}(N^{1/2})$ out of stationarity and $\mathcal{O}(N^{1/3})$ in its stationary regime. Hence, in virtually any situation of practical relevance where the target distribution has a finite number of dimension and/or the algorithm is started out of stationarity, the MALA with annealed proposals turns out to be superior to MALA, and as easily implemented/tuned as the latter.

MSC2020 subject classifications: Primary 60J22; secondary 62F15.

Keywords: Markov chain Monte Carlo, Metropolis-adjusted Langevin algorithm, diffusion limit, optimal scaling, transience

Funding: This work has been supported by the *Natural Sciences and Engineering Research Council of Canada*.

Acknowledgements: The author wishes to thank Michela Ottobre (*U. of Edinburgh*) for interesting exchanges and the subject of this paper.

[*] Université de Montréal, (Département de mathématiques et de statistique), Montréal, (Québec), Canada. Corresponding author (email: bedard@dms.umontreal.ca)

1 Introduction

The Metropolis-Hastings sampler, a device of choice among the popular Markov chain Monte Carlo methods, is used to obtain samples from complicated probability distributions Π (Metropolis et al., 1953; Hastings, 1970). In its elemental form, it builds a Markov process with invariant distribution Π on a state space $\mathcal{S}$ by producing candidates for the Markov chain that are either deemed suitable as the next state of the chain or simply rejected.

Let x_0 be the initial value for the process, either fixed or drawn from an arbitrary distribution μ, and let π be the N-dimensional target density arising from Π with respect to Lebesgue measure. Then, at iteration $k+1$, the Metropolis-Hastings (MH) sampler generates a candidate $y_{k+1} = y$ from a proposal distribution $Q(x_k, y)$ with density $q(x_k, y)$. This candidate is accepted as the next state x_{k+1} of the Markov process with probability $\alpha(x_k, y) = \min\{1, \frac{\pi(y)q(y,x_k)}{\pi(x_k)q(x_k,y)}\}$, otherwise we set $x_{k+1} = x_k$ and the process remains at the current state for another time interval.

The Markov chain produced through this mechanism is reversible with respect to Π, which in turn implies that this target distribution is stationary for the chain. Different choices of proposal distributions Q lead to various performances in terms of efficiency and computational cost. Informed proposal distributions use characteristics of the target Π to produce quality candidates which, on average, turn out to be accepted in a greater proportion than candidates coming from blinded proposal distributions (i.e. distributions that use no such information from Π). For instance, the Metropolis-adjusted Langevin algorithm (MALA) uses the gradient of the target's log-density in order to drag the proposal distribution towards regions of higher target density. The MALA is a Metropolis-Hastings algorithm with (informed) proposal distribution

$$y_{k+1} \quad \sim \quad \mathcal{N}\left(x_k + \delta \nabla \log\left\{\pi(x_k)\right\}, 2\delta I_N\right), \tag{1}$$

with $\delta > 0$ for tuning and where I_N is the $N \times N$ identity matrix.

The proposal distribution of the MALA stems from the discretization of the Langevin diffusion process with time step δ. The invariant distribution of this diffusion process being Π, it makes for an ideal starting point for deriving a good proposal distribution. Because the discretization destroys the invariance of Π, it becomes necessary to compensate via the accept-reject criterion, which is included to preserve detailed balanced with respect to Π. The smaller the time step, the less correction is required from the acceptance probability; time steps that are too small however lead to a slow-moving process.

Besides our choice of proposal distribution Q, the tuning of this proposal also has an impact on the efficiency of the sampler or, in other words, on the speed at which the process explores its state space. Large tunings induce moves that are far from the current state and thus tend to be frequently rejected, leading to a process that stagnates (does not move often enough). Small tunings produce candidates that are close to the current position; although generally accepted, these moves result in a overlong (exploration of the state space. In both cases, the chain exhibits slow mixing, a behaviour that worsens as the dimension N grows.

We thus aim at using a tuning parameter that strikes a balance between these two scenarios. To account for the dimensionality of the target Π, the optimal size of the proposed moves generally is a function of N, i.e. is proportional to $N^{-\zeta}$, for some $\zeta > 0$. With complex and high-dimensional statistical models being ubiquitous nowadays, documenting the efficiency and computational cost of samplers has become essential. Hereafter, we use the expected square jumping distance (ESJD) for measuring the efficiency, or computational cost, of the process. We will say that the cost of the sampler in terms of ESJD is $\mathcal{O}(N^{\zeta})$ if the proposal tuning scales according to $N^{-\zeta}$.

Given its popularity and simplicity of usage, the MALA sampler has been studied extensively and its behaviour is well-documented in the literature. Under certain regularity conditions on the target density, it has been proven that optimally tuned versions of the MALA accept 57.4%, with a computational cost of $\mathcal{O}(N^{1/3})$ in their stationarity phase (Roberts and Rosenthal, 1998). This is much better than the random walk Metropolis (RWM) algorithm with a blinded $\mathcal{N}(x_k, \sigma^2 I_N)$ proposal, which accepts 23.4% of candidates and explores its state space in $\mathcal{O}(N)$ iterations when optimally tuned (Roberts et al., 1997). Regrettably however, the behaviour of MALA samplers in their non-stationary phase is notoriously erratic. To avoid jeopardizing the sampler's convergence, one needs to shrink the tuning parameter, leading to a computational cost of $\mathcal{O}(N^{1/2})$ while in transience. By contrast, the RWM displays a steadier behaviour with its computational cost of $\mathcal{O}(N)$ both in and out of stationarity (Christensen et al., 2005).

More specifically, under certain regularity assumptions on the target distribution, it can be shown that RWM and MALA weakly converge (as $N \uparrow \infty$) towards N-dimensional Langevin diffusion processes. This shed light on one downside of MALA: its proposal distribution, itself originating from a Langevin diffusion process, has been naturally designed to be efficient in high-dimensional, stationary settings. In contexts of practical significance, that is when working with finite-dimensional targets and dealing with samplers that have not yet attained their stationary phase, it might be preferable to select an alternative proposal distribution.

The Metropolis-adjusted Langevin algorithm with annealed proposals (aMALA) introduced in Boisvert-Beaudry and Bédard (2022) is a generalization of MALA. Its proposal distribution is expressed as

$$y_{k+1} \quad \sim \quad \mathcal{N}\left(x_k + \gamma\delta\nabla \log\left\{\pi(x_k)\right\}, 2\delta I_N\right),$$

with $\delta > 0$ and $\gamma \in [1, 2]$. This sampler has been derived using the local- and global-balance properties of Zanella (2020); it is similar to MALA, but features an extra tuning parameter γ in front of the gradient term that may be adjusted to account for the dimension of the target Π. In short, this extra parameter adds flexibility to the proposal distribution by allowing us to increase the weight of the gradient term in the proposal mean. Boisvert-Beaudry and Bédard (2022) argue that γ-values close to 2 and 1 should be used in low- and high-dimensional settings, respectively, with intermediate values ideal for moderate-dimensional targets. We note that MALA, which corresponds to $\gamma = 1$, should be optimal in infinite-dimensional settings; in practice however, the authors could not reach a dimension N large enough to observe this. It would thus appear that MALA is overly conservative when it comes to the biasing of its proposal mean. This is unsurprising, given that Langevin diffusion processes are the building blocks of MALA, while simultaneously describing the asymptotic behaviour of MH samplers as $\delta \downarrow 0$. In finite dimensions, one should therefore assign more weight to the informed portion of the proposal distribution.

To facilitate the tuning of the interpolation parameter γ, the authors suggest using the guideline $\gamma = 1 + 1/N^{1/3}$ and illustrate through various examples that this rule yields a sampler that is nearly optimal in the various examples studied. Letting $\delta = \ell^2/N^{1/3}$ (with $\ell > 0$), $\gamma = 1 + 1/N^{1/3}$, and studying a target Π whose density satisfies the regularity conditions in Roberts and Rosenthal (1998), they proved that that the computational cost of the aMALA is $\mathcal{O}(N^{1/3})$ in its stationary phase. This was to be expected, as the aMALA with $\gamma = 1 + 1/N^{1/3}$ becomes the usual MALA in the limit (since $\gamma \to 1$ as $N \uparrow \infty$).

Now, since γ provides a dimension-specific adjustment of the proposal distribution, we expect this parameter to have a stabilizing effect on the erratic behavior of MALA outside of stationarity. The primary contribution of this paper consists in asymptotic, out-of-stationarity diffusion results for the aMALA sampler, in the context of general non-product target measures. These target measures are defined using a density that is expressed with respect to a Gaussian random field. To this end, we heavily rely on the analysis expounded in Kuntz et al. (2018), where the MALA sampler is used on non-product target densities and started out of stationarity (it is thus studied in its transient phase). We show that the tuning used to obtain our diffusion limit for the aMALA corresponds to an $\mathcal{O}(N^{1/3})$ computational cost, meaning that the behaviour of the sampler remains stable/steady both in and out of stationarity. We also provide some results about the asymptotic behaviour of the aMALA in its stationary phase, along with some guidelines for tuning this sampler. We argue that the computational cost of the stationary aMALA may be as small as $\mathcal{O}(N^{1/5})$ in some contexts. This finding is in line with our previous conclusions, i.e. that aMALA is superior to MALA in terms of computational cost.

In order to present our theoretical results about the aMALA (both in and out of stationarity), we start by introducing the framework in §2. Building on these foundations we present, in §3, the main theoretical results for the aMALA along with an analysis and some tuning guidelines. We then support these findings with numerical explorations and simulation studies in §4. We conclude with an application on Scots pine saplings data in §5 and with a discussion in §6. Proofs are deferred to Appendices **??** to **??**.

2 Framework

2.1 Hilbert space and projections

We use the framework of Kuntz et al. (2018) and work with finite-dimensional target measures that arise from approximations of a measure π on an real separable infinite-dimensional Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle, \|\cdot\|)$, with the canonical norm induced by the inner product. We let the target measure π on $\mathcal{H}$ be defined as

$$\frac{\mathrm{d}\pi}{\mathrm{d}\pi_0} \propto \exp(-\Psi) , \tag{2}$$

with $\pi_0 \sim \mathcal{N}(0, \mathcal{C})$. The measure π is thus absolutely continuous with respect to a Gaussian measure π_0 that has mean 0 and covariance $\mathcal{C}$; furthermore, $\Psi : \tilde{\mathcal{H}} \to \mathbb{R}$ is a real-valued function with $\tilde{\mathcal{H}} \subseteq \mathcal{H}$. The covariance $\mathcal{C}$ is a positive, self-adjoint, trace class operator on $\mathcal{H}$ with eigenbasis $\{\lambda_j^2, \phi_j\}_{j\geq 1}$; we then have

$$\mathcal{C}\phi_j = \lambda_j^2 \phi_j, \quad \forall j \in \mathbb{N} ,$$

and we assume that $\{\phi_j\}_{j\in\mathbb{N}}$ forms a complete orthonormal basis of $\mathcal{H}$.

As MALA and aMALA have both been designed to sample from finite-dimensional distributions of interest, we now let π^N refer to the finite-dimensional projections of the measure π in (2) on the space spanned by the first N eigenvectors of the covariance operator,

$$\mathcal{X}^N := \operatorname{span}\{\phi_j\}_{j=1}^N \subset \mathcal{H} .$$

This basically means that $x^N = \mathcal{P}^N(x) := \sum_{j=1}^N x_j \phi_j$ is the projection of a point $x \in \mathcal{H}$ onto the space $\mathcal{X}^N$, where $x_j = \langle x, \phi_j \rangle$ is the jth component of x and $x \in \mathcal{H}$ has the representation $x = \sum_{j\geq 1} x_j \phi_j$. The jth component of the vector $x^N \in \mathcal{X}^N$ is then expressed as $x_j^N = \langle x^N, \phi_j \rangle$; naturally, $x_j = x_j^N$ for $1 \leq j \leq N$.

Similar notation is used for other vectors and their components, for instance the candidate vector y^N. We also denote

$$\Psi^N := \Psi \circ \mathcal{P}^N \qquad \text{and} \qquad \mathcal{C}_N := \mathcal{P}^N \circ \mathcal{C} \circ \mathcal{P}^N ,$$

meaning that $\mathcal{C}^N$ is an $N \times N$ diagonal matrix with λ_j^2 as its jth diagonal element. According to this notation, our finite-dimensional target measure π^N on $\mathcal{X}^N$ (a space isomorphic to $\mathbb{R}^N$) satisfies

$$\frac{\mathrm{d}\pi^N}{\mathrm{d}\pi_0^N}(x) = M_{\Psi^N} \exp\{-\Psi^N(x)\} , \tag{3}$$

with $\pi_0^N \sim \mathcal{N}(0, \mathcal{C}_N)$ and M_{Ψ^N} a normalization constant. As stated in Kuntz et al. (2018), the sequence of measures $\{\pi^N\}_{N\in\mathbb{N}}$ then converges to the measure π in the Hellinger metric.

2.2 Sobolev-like spaces

We now provide more details about the space $\tilde{\mathcal{H}}$, and identify it in terms of an appropriate Sobolev-like subspace of $\mathcal{H}$.

Hereafter, we denote Sobolev-like subspaces by $\mathcal{H}^s$, $s \in \mathbb{R}$. These subspaces are defined with the following inner products and norms

$$\langle x, y\rangle_s = \sum_{j=1}^{\infty} j^{2s} x_j y_j \qquad \text{and} \qquad \|x\|_s^2 = \sum_{j=1}^{\infty} j^{2s} |x_j|^2 \ .$$

We emphasize the fact that $(\mathcal{H}^s, \langle \cdot, \cdot \rangle_s)$ is a Hilbert space; since $\mathcal{H}^0 = \mathcal{H}$, it directly follows that $\mathcal{H}^s \subset \mathcal{H} \subset \mathcal{H}^{-s}$ for any $s > 0$.

The proof of our result requires that we alternate between some of the spaces just mentioned. To this end, the following elementary inequality will reveal useful

$$|\langle x, y\rangle|^2 \ = \ \left| \sum_{j=1}^{\infty} (j^s x_j)(j^{-s} y_j) \right|^2 \ \leq \ \|x\|_s^2 \, \|y\|_{-s}^2 \ , \quad \forall x \in \mathcal{H}^s, \ y \in \mathcal{H}^{-s} \ . \tag{4}$$

As per the arguments in Kuntz et al. (2018) and Kuntz et al. (2019), one can define an operator that lets us alternate between the Hilbert space $\mathcal{H}$ and the interpolation spaces $\mathcal{H}^s$. This allows us deducing that $\{\hat{\phi}_j = j^{-s}\phi_j\}_{j\geq 1}$ forms an orthonormal basis for $\mathcal{H}^s$; we refer the reader to the above-mentioned articles for more details. Given a random variable $y \sim \mathcal{N}(0, \mathcal{C})$, this implies that we can either express y as

$$y = \sum_{j=1}^{\infty} \lambda_j \rho_j \phi_j \ ,$$

with $\rho_j \sim \mathcal{N}(0,1)$ i.i.d. for $j \geq 1$ or, if $\sum_j \lambda_j^2 j^{2s} < \infty$, as

$$y = \sum_{j=1}^{\infty} (\lambda_j j^s) \rho_j \hat{\phi}_j \ ,$$

again with $\rho_j \sim \mathcal{N}(0,1)$ i.i.d. for $j \geq 1$. We can therefore see y as a mean zero Gaussian random variable with covariance operator $\mathcal{C}$ in $\mathcal{H}$, or with covariance $\mathcal{C}_s$ in $\mathcal{H}s$, where $\mathcal{C}_s$ is diagonal in the basis $\{\hat{\phi}_j\}_{j\geq 1}$, with diagonal entries $j^{2s}\lambda_j^2$.

2.3 Algorithm

The MALA sampler with annealed proposals used to sample from the measure π^N generates candidates according to the following rule

$$y_{k+1}^N \ = \ x_k^N + \gamma\delta\mathcal{C}_N \nabla \log \pi^N(x_k^N) + \sqrt{2\delta}\mathcal{C}_N^{1/2} \xi_{k+1}^N \ , \tag{5}$$

with $\delta > 0$ and $\gamma \in [1, 2]$ as tuning parameters, and with $\xi_{k+1}^N = \sum_{j=1}^N \xi_{j,k+1}\phi_j$, where $\xi_{j,k+1} \sim \mathcal{N}(0,1)$ i.i.d. for $j = 1, \ldots, N$. The inclusion of the matrix $\mathcal{C}_N$ in the proposal kernel leads to a sort of preconditioned aMALA sampler that is similar to the preconditioned

MALA, but with an extra tuning parameter γ. Using (2) and developing the gradient of the log-target, we may reexpress the candidate y_{k+1}^N as

$$y_{k+1}^N \quad = \quad x_k^N - \gamma\delta \left\{ x_k^N + \mathcal{C}_N \nabla \Psi^N(x_k^N) \right\} + \sqrt{2\delta}\mathcal{C}_N^{1/2} \xi_{k+1}^N \ .$$

For any $x^N, y^N \in \mathcal{X}^N$, the proposal design in (5) corresponds to a density

$$q^N(x^N, y^N) \quad \propto \quad \exp\left\{ -\frac{1}{4\delta} \left\| y^N - x^N - \gamma\delta \mathcal{C}_N \nabla \log \pi^N(x^N) \right\|_{\mathcal{C}_N}^2 \right\} \ , \tag{6}$$

where $\| \cdot \|_{\mathcal{C}_N}$ defines a Hilbert-Schmidt norm on $\mathcal{X}^N$, that is

$$\|x^N\|_{\mathcal{C}_N}^2 \quad = \quad \sum_{j=1}^N \frac{1}{\lambda_j^2} |\langle x^N, \phi_j \rangle|^2 \quad = \quad \sum_{j=1}^N \frac{|x_j^N|^2}{\lambda_j^2} \ . \tag{7}$$

Such a norm is induced by the scalar product $\langle \cdot, \cdot \rangle_{\mathcal{C}_N}$ defined as

$$\langle x^N, y^N \rangle_{\mathcal{C}_N} = \langle \mathcal{C}_N^{-1/2} x^N, \mathcal{C}_N^{-1/2} y^N \rangle \ , \quad x^N, y^N \in \mathcal{X}^N \ .$$

Note that a similar norm and scalar product can be defined on $\mathcal{H}$ with the covariance operator $\mathcal{C}$.

Now, the candidate y_{k+1}^N is accepted with probability

$$\alpha^N(x_k^N, y_{k+1}^N) = 1 \wedge \frac{\pi^N(y_{k+1}^N) q^N(y_{k+1}^N, x_k^N)}{\pi^N(x_k^N) q^N(x_k^N, y_{k+1}^N)} \ , \tag{8}$$

where $1 \wedge a = \min\{1, a\}$. If the candidate is accepted, then $x_{k+1}^N = y_{k+1}^N$; otherwise, the process remains where it was, that is $x_{k+1}^N = x_k^N$. The aMALA sampler thus satisfies

$$x_{k+1}^N = \beta_{k+1}^N y_{k+1}^N + (1 - \beta_{k+1}^N) x_k^N \ , \quad x_0^N = \mathcal{P}^N(x_0) \ ,$$

where

$$\beta_{k+1}^N \sim \text{Bernoulli}(\alpha^N(x_k^N, y_{k+1}^N)) \ .$$

Some care should be taken when tuning the parameters γ and δ, which should decrease with the dimension N of the state space $\mathcal{X}^N$. More specifically, let $\delta = \ell_1^2/N^{\zeta_1}$ and $\gamma = 1 + \ell_2^2/N^{\zeta_2}$, where $\ell_1, \ell_2, \zeta_1, \zeta_2$ are positive parameters. As explained in the introduction, if ζ_1 is too large, then δ is small and the sampler only takes tiny steps; at the opposite, if ζ_1 is too small, then proposed steps are overly aggressive and often rejected. Similarly, if ζ_2 is too large on the one hand, then γ approaches 1 too rapidly and the process swiftly behaves like MALA, which is suboptimal in finite-dimensional settings. On the other hand, if ζ_2 is too small, then the biasing term of the proposal mean carries too much weight, leading to candidates that are frequently rejected. It has been showed in Boisvert-Beaudry and Bédard (2022) that the optimal choice in the stationary phase is $\zeta_1 = \zeta_2 = 1/3$, leading to $\delta = \ell_1^2/N^{1/3}$ and $\gamma = 1 + \ell_2^2/N^{1/3}$. In this paper, we show that these same tunings remain optimal in the transient phase of the aMALA. This is an improvement over MALA, which requires an $\mathcal{O}(N^{-1/2})$ out-of-stationarity tuning to preserve its convergence properties. We also show that the computational cost of the aMALA is, in some contexts, as small as $\mathcal{O}(N^{1/5})$ in its stationary phase (in which case we much set $\zeta_1 = \zeta_2 = 1/5$ to study the asymptotic process).

To gain more information about the aMALA sampler in transience, we focus on the sequence $\{S_k^N\}_{k\in\mathbb{N}}$ defined as

$$S_k^N \;=\; \frac{1}{N}\|x_k^N\|_{\mathcal{C}_N}^2 \;=\; \frac{1}{N}\sum_{j=1}^{N}\frac{|x_{j,k}^N|^2}{\lambda_j^2}\ .$$

We note in passing that x_k^N refers to the kth time value of the process, x_j^N refers to the jth component of the vector x^N, and $x_{j,k}^N$ denotes the jth component of the time-k state of the process. Despite our usage of lower-case letters uniformly throughout the paper, we understand that quantities up to time k (and later up to time t) are known, while quantities from time $k+1$ (or after time t) are random variables. The sequence $\{S_k^N\}_{k\in\mathbb{N}}$ is generally not a Markov process; nonetheless, when the aMALA is started at an initial value $x_0^N \in \tilde{\mathcal{H}}$ such that S_0^N is finite, it is well-defined. To study the limiting behaviour of this sequence, we use its continuous interpolant

$$S_t^N = (N^{1/3}t - k)S_{k+1}^N + (k + 1 - N^{1/3}t)S_k^N\ , \quad \frac{k}{N^{1/3}} \le t < \frac{k+1}{N^{1/3}}\ . \tag{9}$$

We denote by $C([0,T],\tilde{\mathcal{H}})$ the space of $\tilde{\mathcal{H}}$-valued functions on $[0,T]$ endowed with the uniform topology. This shall be useful as we will soon present weak convergence results in $C([0,T],\tilde{\mathcal{H}})$ (as $N \uparrow \infty$) for the continuous interpolant just introduced.

When the dimension N is fixed, the aMALA process lives on $\mathcal{X}^N$ and has invariant distribution π^N. However, considering that we want to study the scaling limit of this process as $N \uparrow \infty$, it will be preferable to carry out the analysis in $\mathcal{H}$. The first N components of the $\mathcal{H}$-valued vector we study coincide with x^N, while the remaining components are not updated and stay at their initial state. Specifically, this process may be written in component-wise notation as

$$\begin{aligned} x_{j,k+1} \;&=\; x_{j,k+1}^N \\ &=\; x_{j,k}^N + \beta_{k+1}^N\left[\frac{\ell_1^2}{N^{1/3}}\left(1 + \frac{\ell_2^2}{N^{1/3}}\right)[\mathcal{C}_N\nabla\log\pi^N(x_k^N)]_j + \sqrt{\frac{2\ell_1^2}{N^{1/3}}}\lambda_j\xi_{j,k+1}^N\right]\ , \end{aligned}$$

$\forall j \le N$, and $x_{j,k+1} = x_{j,k} = 0$, $\forall j \ge N+1$.

2.4 Assumptions

We now introduce some assumptions on the function Ψ and the covariance operator $\mathcal{C}$ of the measure π_0, both appearing in (2). We first suppose that there exists some fixed $s \geq 0$ such that $\Psi : \mathcal{H}^s \to \mathbb{R}$; simply put, this means that we let $\tilde{\mathcal{H}} = \mathcal{H}^s$.

For each $x \in \mathcal{H}^s$, the gradient $\nabla\Psi(x)$ is an element of the dual $\mathcal{L}(\mathcal{H}^s, \mathbb{R})$ of $\mathcal{H}^s$, containing the linear functionals on $\mathcal{H}^s$. We emphasize that for $x \in \mathcal{H}^s$, we may identify $\mathcal{L}(\mathcal{H}^s, \mathbb{R}) = \mathcal{H}^{-s}$ and view the gradient $\nabla\Psi(x)$ as an element of $\mathcal{H}^{-s}$. We then impose the following conditions on Ψ and $\mathcal{C}$:

1. **Decay of eigenvalues λ_j^2 of $\mathcal{C}$:** There exists a constant $\kappa > s + \frac{1}{2}$ such that
$$j^{-\kappa} \lesssim \lambda_j \lesssim j^{-\kappa} \ ,$$
where $a_j \lesssim b_j$ means that there is a constant $K > 0$ (independent of j) such that $a_j < Kb_j$ for all j.

2. **Domain of Ψ:** The functional Ψ is defined everywhere on $\mathcal{H}^s$.

3. **Derivatives of Ψ:** The gradient of $\Psi(x)$ is bounded and globally Lipschitz:
$$\|\nabla\Psi(x)\|_{-s} \lesssim 1 \ , \quad \|\nabla\Psi(x) - \nabla\Psi(y)\|_{-s} \lesssim \|x - y\|_s \ .$$

The decay of eigenvalues in the first assumption ensures that the trace of $\mathcal{C}_s$ in $\mathcal{H}^s$ is finite, i.e. $\mathrm{Trace}_{\mathcal{H}^s}(\mathcal{C}_s) = \sum_{j=1}^{\infty} \lambda_j^2 j^{2s} < \infty$, so that $\pi_0(\mathcal{H}^s) = 1$. Furthermore, we note that since $\lambda_j j^s \downarrow 0$ as $j \uparrow \infty$, then the sequence $\{\lambda_j j^s\}_j$ is bounded, hence $\lambda_j j^s \leq C$ for some constant $C > 0$.

The above assumptions have useful consequences. Before proceeding with our main result in the next section, we present two lemmas that will be handy in our specific context; the proof of these results may be found in Appendix A of Kuntz et al. (2018).

Lemma 1 (Lemma 2.1 of Kuntz et al. (2018))**.** *Suppose that the previous assumptions hold. Then,*

1. *The function $\mathcal{C}\nabla\Psi(x)$ is bounded and globally Lipschitz on $\mathcal{H}^s$, that is*
$$\|\mathcal{C}\nabla\Psi(x)\|_s \lesssim 1 \qquad \text{and} \qquad \|\mathcal{C}\nabla\Psi(x) - \mathcal{C}\nabla\Psi(y)\|_s \lesssim \|x - y\|_s \ .$$
Therefore, the function $F(z) = -z - \mathcal{C}\nabla\Psi(z)$ satisfies
$$\|F(x) - F(y)\|_s \lesssim \|x - y\|_s \qquad \text{and} \qquad \|F(x)\|_s \lesssim 1 + \|x\|_s \ .$$

2. *The function $\Psi(x)$ is globally Lipschitz and therefore $\Psi^N(x) = \Psi(\mathcal{P}^N(x))$ also is globally Lipschitz:*
$$\left|\Psi^N(y) - \Psi^N(x)\right| \lesssim \|y - x\|_s \ .$$

Lemma 2 (Lemma 2.2 of Kuntz et al. (2018))**.** *Suppose that the previous assumptions hold. Then, the following holds for the function Ψ^N and its gradient $\nabla\Psi^N = \nabla(\Psi \circ \mathcal{P}^N)$:*

1. *If the bounds specified in the third assumption hold for Ψ, then they hold for Ψ^N as well:*

$$\left\|\nabla\Psi^N(x)\right\|_{-s} \lesssim 1 \ , \qquad \left\|\nabla\Psi^N(x) - \nabla\Psi^N(y)\right\|_{-s} \lesssim \|x-y\|_s \ .$$

2. *Moreover,*

$$\left\|\mathcal{C}_N\nabla\Psi^N(x)\right\|_s \lesssim 1 \quad \text{and} \quad \left\|\mathcal{C}_N\nabla\Psi^N(x)\right\|_{\mathcal{C}_N} \lesssim 1 \ .$$

3 Asymptotic tuning results for aMALA

3.1 Out-of-stationarity tuning results

With the above framework in place, we are now ready to state our main result about the asymptotic, out-of-stationarity behaviour of the MALA with annealed proposals.

Define $\mathcal{H}^s_\cap$ to be the set of $\mathcal{H}^s$-values x whose $\mathcal{C}_N$-norm squared, divided by N, remains finite when $N \uparrow \infty$. In other words,

$$\mathcal{H}^s_\cap \quad = \quad \left\{ x \in \mathcal{H}^s : \lim_{N\uparrow\infty} \frac{1}{N}\sum_{j=1}^{N} \frac{|x_j|^2}{\lambda_j^2} < \infty \right\} \ .$$

For initial values chosen in this set, the following result holds.

Theorem 1. *Consider a target measure π as in* (2)*, for which the assumptions in Section 2.4 hold. Let $\{x_k^N\}_{k\in\mathbb{N}}$ be the Metropolis-Hastings algorithm that samples from π using the proposal design of the aMALA in* (5)*, with $\delta = \ell_1^2/N^{1/3}$, $\gamma = 1 + \ell_2^2/N^{1/3}$, and $\ell_1^2 = 2\ell_2^2$.*

Then, for $T > 0$ and any deterministic initial datum $x_0^N = \mathcal{P}^N(x_0)$, where x_0 is any point in $\mathcal{H}^s_\cap$, the continuous interpolant S_t^N of the sequence $\{S_k^N\}_{k\in\mathbb{N}} \subseteq \mathbb{R}_+$ (defined in (9)) converges in probability in $\mathcal{C}([0,T];\mathbb{R})$ to $S_t \in \mathbb{R}_+ := \{s \in \mathbb{R} : s \geq 0\}$, which is the solution of the ordinary differential equation (ODE)

$$\begin{aligned} \mathrm{d}S_t &= 2\ell_1^2(1-S_t)\left\{1 \wedge e^{\ell_1^4\ell_2^2(S_t-1)}\right\}\mathrm{d}t \ , \qquad (10) \\ S_0 &= \lim_{N\uparrow\infty} S_0^N = \lim_{N\uparrow\infty}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{\lambda_j^2}\left|x_{j,0}^N\right|^2 \ . \end{aligned}$$

Proof. The proof of this result may be found in Appendices **??** to **??**. □

The above theorem assumes that the initial datum of the chains $\{x_k^N\}$ is assigned deterministically. Nevertheless, as discussed in Kuntz et al. (2018), the same statement holds for random initial data, as long as the process is started out of stationarity (i.e. x_0^N is not drawn from π^N, or from any change of measure from π^N), and S_0^N has bounded (uniformly in N) moments of sufficiently high order that are independent of any other source of noise in the algorithm.

Since the time step implied by the interpolation S_t^N is $N^{-1/3}$, Theorem 1 implies that the number of iterations required by the Markov process in its transient phase is $\mathcal{O}(N^{1/3})$. This is in contrast to MALA, which requires $\mathcal{O}(N^{1/2})$ steps out of stationarity; indeed, any tuning parameter larger than $\ell^2/\sqrt{N}$ for $\ell^2 > 0$ leads to a sampler that behaves erratically. In particular, depending on its initial state, MALA either accepts every move or rejects all of them.

The parameter γ in the proposal mean of the aMALA then appears to stabilize the sampler's behaviour while in transience. In particular, attributing a larger weight to the biasing term of the proposal mean allows using a more aggressive step size δ, even when starting the sampler out in the tails (as long as the weight decreases towards 1 at a predetermined rate as $N \uparrow \infty$). By comparison, the unsuitability of the bias in MALA's proposal mean forces users to rely on a tuning δ that is more conservative in order to hope reaching stationarity – eventually. This is thus a significant improvement over MALA.

The condition $\ell_1^2 = 2\ell_2^2$ is of the utmost importance when setting $\zeta_1 = \zeta_2 = 1/3$; it is required to eliminate the terms that are $\mathcal{O}(N^{1/3})$ in the acceptance probability (see Appendix **??**). Without this condition, we would face an asymptotic behaviour similar to MALA, i.e. that becomes erratic as N grows, due to a degeneration of the acceptance rate. This behaviour would then need to be corrected by relying on less aggressive tunings of the form $\delta = \ell_1^2/N^{1/2}$ and $\gamma = 1 + \ell_2^2/N^{1/2}$. We note that with $\zeta_1 = \zeta_2 = 1/2$, the adjustment of the tuning parameters ℓ_1^2 and ℓ_2^2 is not as sensitive as with $\zeta_1 = \zeta_2 = 1/3$ (no need to preserve a specific relation between ℓ_1^2 and ℓ_2^2), but convergence is slower as we generate more conservative candidates. Generally speaking, we could also select other combinations of $1/3 < \zeta_1, \zeta_2 < 1/2$, with $\zeta_1 = \zeta_2$ and tunings that satisfy $\ell_1^2 = 2\ell_2^2$, but the asymptotic behaviour would not be as interesting. Specifically, the process would suffer from a slower exploration of its space (in transience and in stationarity, compared to $\zeta_1 = \zeta_2 = 1/3$), as we would settle for a less aggressive (i.e. smaller) pair of tuning parameters.

From Theorem 1, the condition $\ell_1^2 = 2\ell_2^2$ leads to the ODE

$$\mathrm{d}S_t \;=\; 2\ell_1^2(1-S_t)\left\{1 \wedge \mathrm{e}^{\frac{\ell_1^6}{2}(S_t-1)}\right\}\mathrm{d}t\ . \tag{11}$$

How should the constant ℓ_1^2 be chosen? To minimize the time spent in transience, $|b_{\ell_1}(s)|$ should be as large as possible, while being positive for $s < 1$ and negative for $s > 1$. In Figure 1, the function $b_{\ell_1}(s) = 2\ell_1^2(1-s)\{1 \wedge \mathrm{e}^{\ell_1^6(s-1)/2}\}$ is plotted against s for different choices of $\ell_1 > 0$. Among those choices, we tested $\ell_1^2 = 1.36$, which is the optimal tuning for

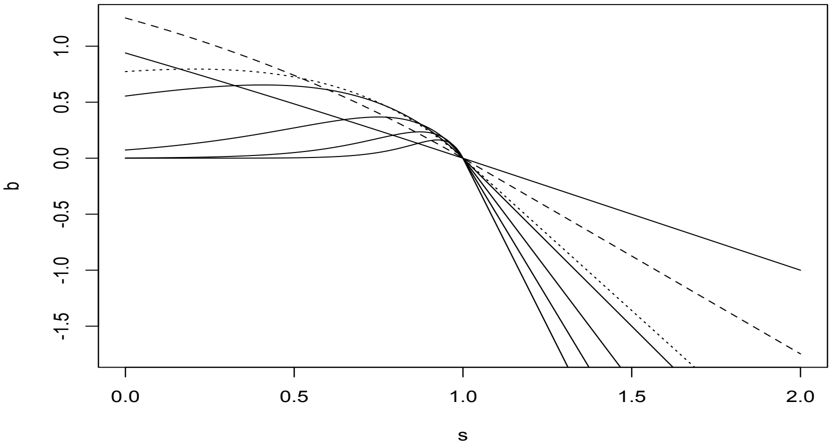

Figure 1: A collection of the $b_{\ell_1}(\cdot)$ functions in (11) for various values of ℓ_1^2. The dashed curve is for $\ell_1^2 = (2/3)^{1/3}$, the dotted curve for $\ell_1^2 = 1.36$, and the solid curves are for $\ell_1^2 = 0.5, 1.5, 2, 2.5, 3$.

a stationary MALA implemented on a high-dimensional standard normal target distribution; $\ell_1^2 = 2$, which corresponds to the simple choice $\ell_2^2 = 1$ discussed in Boisvert-Beaudry and Bédard (2022); and also $\ell_1^2 = (2/3)^{1/3}$, which maximizes the function $b_{\ell_1}(s)$ at $s = 0$. From the graph, it is quite obvious that there is no value of ℓ_1 for which the speed of convergence is uniformly maximized.

A natural approach would be to tune ℓ_1^2 as a function of the initial state of the Markov chain. This would be relatively easy to do in practice, as the optimization problem does not depend on Ψ or $\mathcal{C}$ (i.e. on the particular target distribution π studied). An alternative would consist in using the pair $\ell_1^2 = (2/3)^{1/3}$ and $\ell_2^2 = \ell_1^2/2 = (1/12)^{1/3}$, which maximizes the function $b_{\ell_1}(s)$ at $s = 0$. This choice seems reliable and close to optimality for a broad range of starting values, so it could be used more widely than just with $S_0^N = 0$. As the chain evolves and proceeds towards stationarity, a pragmatic strategy would be to adjust ℓ_1^2 so as to approach a prespecified acceptance rate (the optimal acceptance rate for the specific context under study, for instance). In practice, this could be achieved in an automatic fashion; this shall be perused separately.

3.2 Tuning results under stationarity

3.2.1 aMALA without preconditioning

It has already been established in Boisvert-Beaudry and Bédard (2022) that the computational cost of the aMALA with tunings of the form $\delta = \ell_1^2/N^{1/3}$ and $\gamma = 1 + \ell_2^2/N^{1/3}$ is

$\mathcal{O}(N^{1/3})$. In that paper, target densities of the form $\pi^N(x^N) = \prod_{i=1}^N f(x_i^N)$ were studied, where the one-dimensional density $f(x_i^N)$ satisfied sufficient smoothness conditions. The aMALA sampler studied did not use any preconditioning, meaning that the matrix $\mathcal{C}_N$ in (5) was assumed to be I_N, the $N \times N$ identity matrix, independently of the target density under study.

Under that setting, it can be shown that the asymptotic behaviour of the aMALA sampler corresponds to that of a Langevin diffusion process with speed measure $\upsilon(\ell_1, \ell_2)$,

$$\begin{aligned}\upsilon(\ell_1, \ell_2) \quad = \quad & 2\ell_1^2 \Phi \left(-\frac{1}{2} \left[\ell_1^6 \left\{ \frac{5}{6} \mathbb{E}[\{(\log f(x_i^N))'''\}^2] + \frac{1}{2} \mathbb{E}[\{(\log f(x_i^N)''\}^3] \right\} \right. \right. \\ & \left. \left. -2\ell_1^4 \ell_2^2 \mathbb{E}[\{(\log f(x_i^N)''\}^2] - 2\ell_1^2 \ell_2^4 \mathbb{E}[(\log f(x_i^N)''] \right]^{1/2} \right) \ . \qquad (12)\end{aligned}$$

This function is maximized, with respect to ℓ_2^2, at the unique value

$$\hat{\ell}_2^2 = \frac{\ell_1^2}{2} \frac{\mathbb{E}[\{(\log f(x_i^N)''\}^2]}{\mathbb{E}[\{(\log f(x_i^N)'\}^2]} \ ; \qquad (13)$$

this leads to $\upsilon(\ell_1, \hat{\ell}_2) = 2\ell_1^2 \Phi(-\ell_1^3 K/2)$ with

$$K \quad = \quad \left\{ \frac{5}{6} \mathbb{E}[\{(\log f(x_i^N)'''\}^2] + \frac{1}{2} \mathbb{E}[\{(\log f(x_i^N)''\}^3] \right\}^{1/2} \ ,$$

which is in turn maximized at the unique value $\hat{\ell}_1^2$ for which the acceptance rate is equal to 57.4%. This may be verified by following along Roberts and Rosenthal (1998)'s proof for MALA; in our case however, the limiting acceptance rate of aMALA is impacted by its γ parameter, resulting in two extra terms involving ℓ_2^2 and ℓ_2^4, respectively, in (12). We emphasize that when setting $\ell_2 = 0$, (12) is in agreement with the speed measure found in Theorem 1 of Roberts and Rosenthal (1998); we must however keep in mind that our parameterization for MALA is slightly different from theirs, as they let $h = 2\delta$ in (1) and work in terms of h, which would be the cause of any apparent discrepancy.

Contrarily to what was initially thought, the optimality of the 57.4% acceptance rate does not hold across all choices of ℓ_2^2. For this acceptance rate to be asymptotically optimal, (13) must be satisfied. For other ℓ_2^2 tunings, the $\mathcal{O}(N^{1/3})$ computational cost holds, but the asymptotically optimal acceptance rate varies according to the value of ℓ_2^2 selected and the specific target distribution studied.

Now, in the trivial case where the target distribution is the N-dimensional normal distribution with mean vector μ_N and covariance matrix $\sigma^2 I_N$ ($\sigma^2 > 0$), it turns out that the computational cost of the aMALA just discussed is potentially much cheaper. In fact, if (13) holds, then several simplifications occur and the computational cost drops to $\mathcal{O}(N^{1/5})$. In particular, setting $\zeta_1 = \zeta_2 = 1/5$, using a time step of $N^{-1/5}$ in our interpolated processes, and following along the proof of Roberts and Rosenthal (1998)'s Theorem 1 leads to a limiting Langevin diffusion process as before, but with speed measure

$$\upsilon(\ell_1, \hat{\ell}_2) \;=\; 2\ell_1^2 \Phi\left(-\frac{\ell_1^5}{2\sqrt{2}\;\sigma^5}\right) \,.$$

This speed measure is optimized at the unique value $\hat{\ell}_1^2$ for which the asymptotically optimal acceptance rate is $2\Phi\left(-\hat{\ell}_1^5/\{2\sqrt{2}\;\sigma^5\}\right) = 70.4\%$; this corresponds to $\hat{\ell}_1^2 = 1.0287\sigma^2$ and $\hat{\ell}_2^2 = 0.5143$.

3.2.2 aMALA with preconditioning $\mathcal{C}_N$

The aMALA with proposal distribution (5) displays an asymptotic behaviour in its stationary phase that is slightly different from the aMALA in §3.2.1. In particular, using a preconditioning matrix $\mathcal{C}_N$ in the aMALA of Boisvert-Beaudry and Bédard (2022) allows us sending the trajectory farther away in a given iteration, leading to a much faster exploration of the state space. In fact, we obtain asymptotic results that are similar to the special case of the standard normal target mentioned above. For a target density π as in (2) for which the assumptions in Section 2.4 hold, we find that the computational cost of the sampler started in its stationary phase is $\mathcal{O}(N^{1/5})$. More specifically, setting $\delta = \ell_1^2/N^{1/5}$ and $\gamma = 1 + \ell_2^2/N^{1/5}$ with $\ell_2^2 = \ell_1^2/2$, we find that our sequence of interpolated processes weakly converges (as $N \uparrow \infty$) to a limiting Langevin diffusion process with speed measure

$$\upsilon(\ell_1) \;=\; 2\ell_1^2 \Phi\left(-\frac{\ell_1^5}{2\sqrt{2}}\right) \,. \tag{14}$$

This speed measure is optimized at a unique value $\hat{\ell}_1^2$, and this optimal tuning corresponds to an asymptotically optimal acceptance rate of $2\Phi\left(-\hat{\ell}_1^5/\{2\sqrt{2}\}\right) = 70.4\%$.

Note that if the relation between ℓ_1^2 and ℓ_2^2 were to be violated, we would simply work with a sampler whose computational cost remains equal to $\mathcal{O}(N^{1/3})$ in its stationary phase. In that case, the computational cost would be similar to that of MALA, both in and out of stationarity, and the asymptotically optimal acceptance rate would be a function of ℓ_1^2, ℓ_2^2, and π (through the speed measure of the limiting Langevin diffusion process obtained).

As before, the proof of the above result may be obtained by adjusting Roberts and Rosenthal (1998)'s proof of their Theorem 1 to account for the parameter γ and the preconditioning matrix $\mathcal{C}_N$. Although we do not provide the proofs of the results discussed in §3.2, we explore them through simulation studies in §4. In the next section, we compare efficiency curves arising from different settings of the aMALA with the theoretical curves corresponding to the speed measures just discussed.

Several conclusions may be drawn from the results of §3. First of all, one has access to significant efficiency gains in terms of computational cost by judiciously selecting the parameter ℓ_2^2 in (5) (according to (13) when there is no preconditioning, and $\ell_2^2 = \ell_1^2/2$ when there is). Second of all, it is quite obvious that the use of a preconditioning matrix $\mathcal{C}_N$, when available, leads to significant efficient gains in terms of computational cost. Out of stationarity, this cost remains at $\mathcal{O}(N^{1/3})$ in both scenarios, but in the stationary phase, using a preconditioned version of the aMALA leads to a computational cost as small as $\mathcal{O}(N^{1/5})$. In both cases (preconditioned or not), this represents an improvement over the usual MALA, which offers $\mathcal{O}(N^{1/2})$ and $\mathcal{O}(N^{1/3})$ computational costs in and out of stationarity, respectively. In practice, we usually do not have access to the exact preconditioning matrix $\mathcal{C}_N$. It is reasonable to assume that the computational cost will lie somewhere between $\mathcal{O}(N^{1/3})$ and $\mathcal{O}(N^{1/5})$ in the stationary phase, depending on the accuracy of our estimate of $\mathcal{C}_N$.

4 Simulations

4.1 Normal target distribution

Consider the target density $\pi^N(x^N) \propto \exp\{-\left\|x^N\right\|^2/2\}$, where x^N has dimension $N =$ 1,000. Suppose that we run the aMALA in (5) with $\mathcal{C}_N = I_N$ and initial value 0 to obtain a sample from this target density. We first set $\delta = \ell_1^2/N^{1/3}$ with $\ell_1^2 = (2/3)^{1/3}$ and $\gamma = 1 + \ell_2^2/N^{1/3}$ with $\ell_2^2 = (1/12)^{(1/3)}$, which are the optimal tuning parameters for an out-of-stationarity aMALA started at the origin. The top line of Figure 2 presents trace plots of $\|x\|^2$ under these settings; the left graph illustrates the first 10,000 iterations, while the right one depicts the first 300 iterations of the same output together with the theoretical curve. The latter, in red, is the solution of the ordinary differential equation in (10). The bottom line of Figure 2 displays similar trace plots obtained with MALA; for this, we use $\delta = \ell_1^2/N^{1/2}$ with $\ell_1^2 = 1$ and $\gamma = 1$, which are the optimal parameters for this transient MALA started at the origin (results from Christensen et al. (2005) adjusted to account for our specific parameterization).

As expected from the asymptotic ODE, the initial convergence of aMALA appears to be deterministic for the first 20 iterations. The sampler moves extremely rapidly toward stationarity, and once there, continues mixing quite efficiently. The stationary phase is reached in approximately 25 iterations, which is about the third of the time required by MALA in a similar context (see bottom right graph, which requires about 75 iterations before stabilizing). Using an $\mathcal{O}(N^{-1/3})$ tuning for δ in aMALA thus leads to significant improvements over MALA, which cannot handle tunings larger than $\mathcal{O}(N^{-1/2})$ in its transient phase.

In both cases, the extremely high acceptance rates obtained (97.9% for aMALA and 95.1% for MALA) indicate that the candidates produced are too conservative for the stationary phase of these samplers. Consequently, these algorithms would benefit from an update of their tuning parameters mid-run. Once stationarity is reached, letting $\delta = \ell_1^2/N^{1/5}$ with $\ell_1^2 = 1.03$ and $\gamma = 1 + \ell_2^2/N^{1/5}$ with $\ell_2^2 = \ell_1^2/2$ in the aMALA yields an acceptance

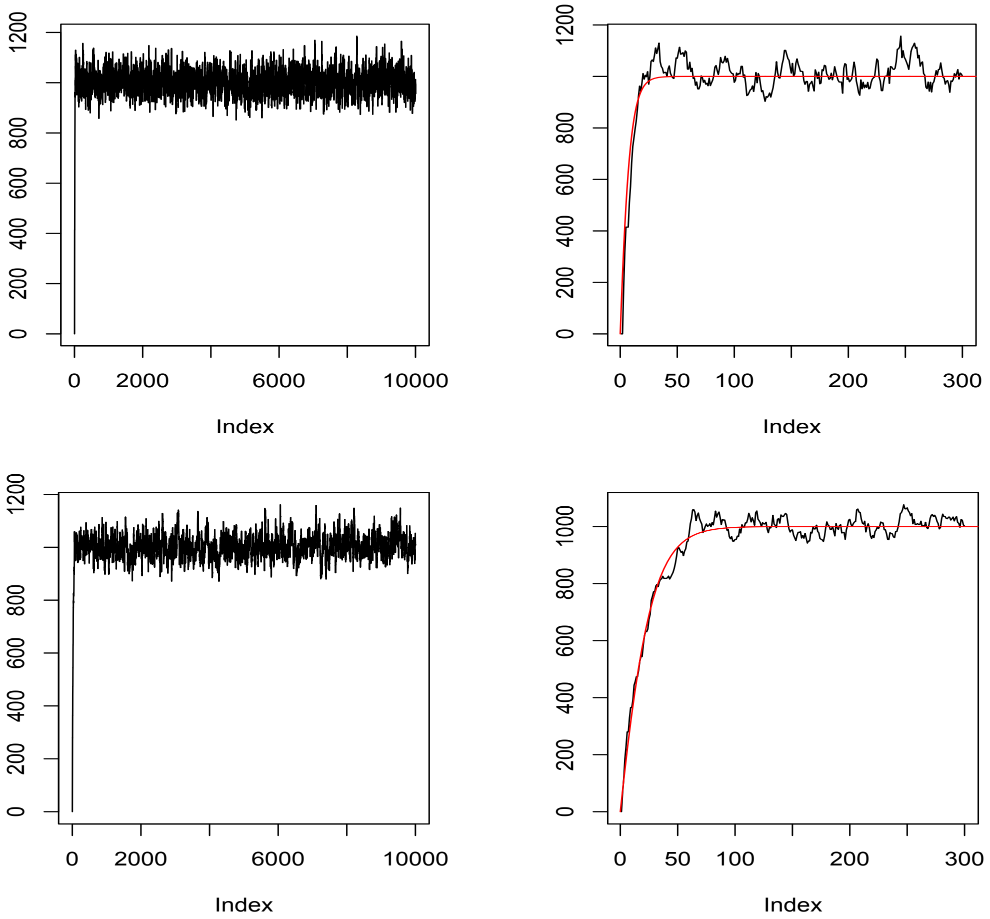

Figure 2: Trace plots of $\|x\|^2$ for simulating a 1,000-dimensional standard normal distribution using aMALA started at the origin. Top: $\delta = \ell_1^2/N^{1/3}$ and $\gamma = 1 + \ell_2^2/N^{1/3}$, with $\ell_1^2 = (2/3)^{1/3}$ and $\ell_2^2 = (1/12)^{1/3}$; bottom: $\delta = \ell_1^2/N^{1/2}$ with $\ell_1^2 = 1$ and $\gamma = 1$ (MALA). The first 10,000 iterations are depicted on the left and the first 300 on the right; also included is a solid red line that represents the solution to $f'(t) = b_{\ell_1}(f(t))$ (top) and to $f'(t) = 2\{1 - f(t)\}(1 \wedge \exp[-\{1 - f(t)\}/2])$ (bottom).

rate that is close to 70.4% hence, according to §3.2, to a nearly optimal version of the algorithm. To validate this claim, we define the average squared jumping distance as

$$\mathrm{ASJD}(x^N) = \frac{1}{KN} \sum_{k=0}^{K-1} \sum_{j=1}^{N} (x_{j,k+1}^N - x_{j,k}^N)^2$$

and use it here as a measure of the efficiency of an algorithm, i.e. a measure of the speed at which the process explores its state space.

Figure 3 pictures the efficiency (ASJD) of aMALA against its acceptance rate. Each point in a given graph is obtained by running 100,000 iterations of an aMALA started at 1 (near stationarity) using a different value of ℓ_1^2. Specifically, we record the ASJD and the proportion of accepted candidates of that run, and then place the corresponding point on the graph. The four graphs presented make different assumptions on the parameter ℓ_2^2. In the top left graph, we use $\mathcal{O}(N^{-1/5})$ tunings and choose ℓ_2^2 such that the condition $\ell_1^2 = 2\ell_2^2$ holds; in the other three graphs, $\mathcal{O}(N^{-1/3})$ tunings are used and ℓ_2^2 is fixed at 0.5/2, 1.36/2, and 1, respectively. We also include, on each graph, the theoretical efficiency curve under stationarity as obtained from $\upsilon(\ell_1)$ in (14) (top left) and from (12) for the other graphs. In our context, the speed measure in (12) reduces to

$$\upsilon(\ell_1, \ell_2) \quad = \quad 2\ell_1^2 \Phi \left(-\frac{1}{2} \left\{ \frac{\ell_1^6}{2} - 2\ell_1^4 \ell_2^2 + 2\ell_1^2 \ell_2^4 \right\}^{1/2} \right) \,,$$

where ℓ_2^2 is respectively fixed at 0.5/2, 1.36/2, and 1.

The curve in the top left graph is maximized at an acceptance rate of 70% and presents the highest ASJD among all four graphs. According to our numerical explorations, letting $\ell_2^2 = \ell_1^2/2$ and tuning aMALA to accept about 70% of candidates in its stationary phase thus leads to an optimal sampler. This agrees with the theoretical results presented in §3.2. Breaking the condition $\ell_1^2 = 2\ell_2^2$ does not appear like a good approach to optimize the exploration of the state space. Indeed, efficiency curves are not as high in those cases and, as ℓ_2^2 grows, the sampler becomes more sensitive to the choice of tuning ℓ_1^2 (for large values of ℓ_2^2, underestimating ℓ_1^2 might lead to a sudden drop in terms of efficiency and acceptance rate). From those graphs, we realize that when δ is too small compared to γ, candidates go through a phase where they are more strongly guided by their proposal mean than by their proposal variance. This leads to an acceptance rate that gradually decreases before going back to 1. Candidates in that region of the efficiency curve lack randomness and become slightly more difficult to accept. However, at some point, δ becomes small enough for the candidates produced to be sufficiently close to the current state of the process so as to be virtually always accepted. The curve obtained using the relation $\ell_1^2 = 2\ell_2^2$ is, by far, the best among the scenarios presented. This comes as no surprise given the results expounded in §3.2.

As can be witnessed from the graphs in Figure 3, it takes a relatively long time before the asymptotics kick in. Even with a target distribution in $N = 1,000$ dimensions, the

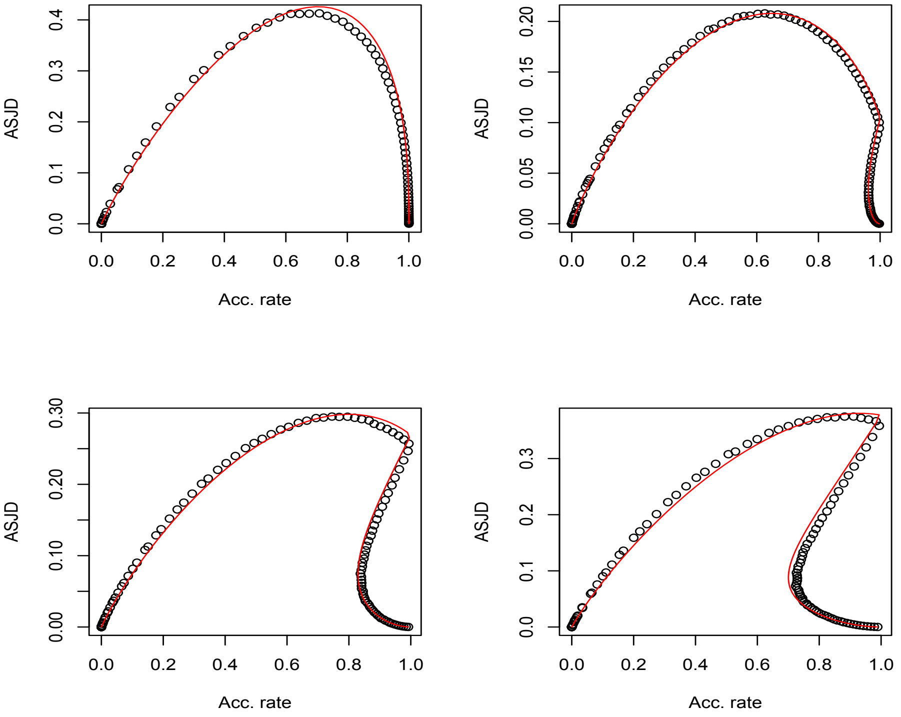

Figure 3: Efficiency against acceptance rate for simulating a $1,000$-dimensional standard normal distribution using 100,000 iterations of aMALA started at $x = (1, \ldots, 1)$. We set $\delta = \ell_1^2/N^\zeta$, $\gamma = 1 + \ell_2^2/N^\zeta$, and use a range of values for ℓ_1^2. Top: $\ell_2^2 = \ell_1^2/2$, $\zeta = 1/5$ (left), $\ell_2^2 = 0.5/2$, $\zeta = 1/3$ (right). Bottom: $\ell_2^2 = 1.36/2$, $\zeta = 1/3$ (left), $\ell_2^2 = 1$, $\zeta = 1/3$ (right). The solid red lines represent the theoretical efficiency curves.

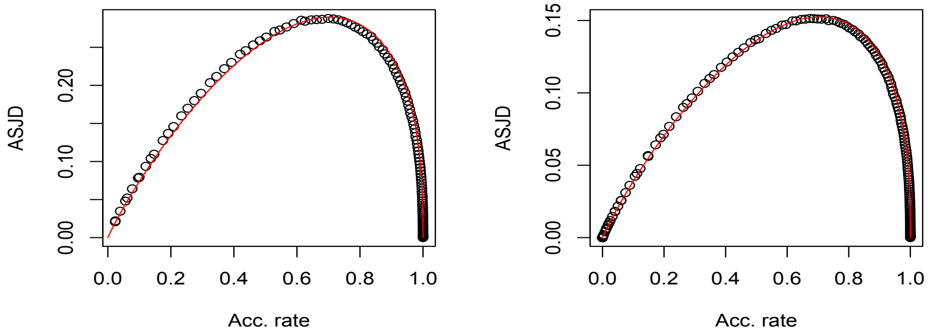

Figure 4: Efficiency against acceptance rate for simulating an N-dimensional standard normal distribution using 100,000 iterations of aMALA started at $x = (1, \ldots, 1)$. We set $\delta = \ell_1^2/N^{1/5}$, $\gamma = 1+\ell_2^2/N^{1/5}$, $\ell_2^2 = \ell_1^2/2$, and use a range of values for ℓ_1^2. Left: $N = 5,000$. Right: $N = 100,000$. The solid red lines represent the theoretical efficiency curves.

efficiency curves of the sampler do not perfectly agree with the theoretical efficiency curves. When increasing N, we however reach a near-perfect agreement between the theoretical curves and the simulated ones. Figure 4 displays graphs that are similar to the top left graph in Figure 3 with $\ell_2^2 = \ell_1^2/2$, but for standard normal target distributions in $N = 5,000$ and $N = 100,000$ dimensions, respectively.

4.2 Beta-logistic target distribution

We now depart from the normality assumption and validate the asymptotic results presented in §3.2.1 when aMALA does not use preconditioning. We study a Beta-logistic target distribution: if $B \sim \text{Beta}(a, b)$ on $(0, 1)$ with $a, b > 0$, then $X = \text{logit}(B) = \log\{B/(1-B)\}$ has a beta-logistic distribution on $\mathbb{R}$ with density

$$f(x) \quad = \quad \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mathrm{e}^{ax}(1+\mathrm{e}^{x})^{-(a+b)} \ , \ x \in \mathbb{R} \ .$$

We consider $N = 1,000$ independent copies of this random variable with parameters $a = b = 2$ and run 100,000 iterations of aMALA with proposal

$$\mathcal{N}\left(x^N + \delta\gamma\left\{a - (a+b)\frac{\exp(x^N)}{1+\exp(x^N)}\right\}, 2\delta I_N\right) \ .$$

The initial state of the sampler is directly drawn from the Beta-logistic$(2, 2)$ so as to focus on the stationary phase of the sampler. Figure 5 presents efficiency graphs of ASJD against acceptance rate for a range of $\delta = \ell_1^2/N^{1/3}$ values, similar to what was done in the previous

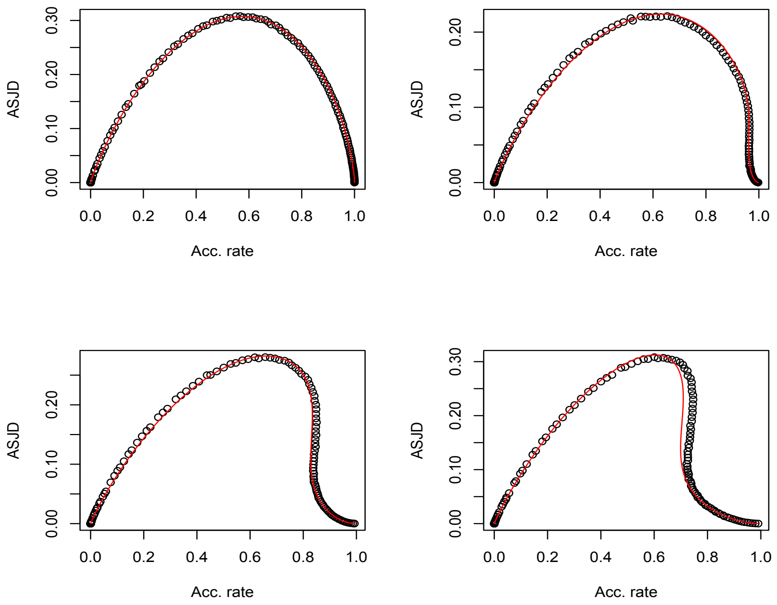

Figure 5: Efficiency against acceptance rate for a 1,000-dimensional Beta-logistic target (100,000 iterations of aMALA in stationarity with $\delta = \ell_1^2/N^{1/3}$, $\gamma = 1 + \ell_2^2/N^{1/3}$, and a range of values for ℓ_1^2). Top: ℓ_2^2 as in (13) (left), $\ell_2^2 = 0.5/2$ (right). Bottom: $\ell_2^2 = 1.36/2$ (left), $\ell_2^2 = 1$ (right). The solid red lines represent the theoretical efficiency curves.

section. As before, each of the four graphs illustrates the behaviour associated to a different $\gamma = 1 + \ell_2^2/N^{1/3}$ value. The solid red lines represent the theoretical efficiency curves of $\upsilon(\ell_1, \ell_2)$ in (12) against the asymptotic acceptance rate, $a(\ell_1, \ell_2) = \upsilon(\ell_1, \ell_2)/\ell_1^2$.

As expected, the optimal version of aMALA arises when (13) holds; in the current context,

$$\hat{\ell}_2^2 \quad = \quad \frac{\ell_1^2}{2}\frac{(a+b+1)(a+b)^2}{ab}\frac{\Gamma(a+b)}{\Gamma(a)}\left(\frac{\Gamma(a+2)}{\Gamma(a+b+2)} - 2\frac{\Gamma(a+3)}{\Gamma(a+b+3)} + \frac{\Gamma(a+4)}{\Gamma(a+b+4)}\right) \ .$$

In this case, the optimal ℓ_1^2 is the value for which the proportion of accepted candidates is 57.4%, as prescribed by the theory. We note that the efficiency curves obtained with a non-preconditioned version of aMALA along with arbitrary ℓ_2^2 values appear smoother than those obtained with some preconditioning. In conclusion, even though we know that asymptotically optimal acceptance rates vary as a function of ℓ_2^2, these rates seem to remain relatively close to 70.4% in the first example, and to 57.4% in the second one. These might serve as useful guidelines in practice.

5 Preconditioned aMALA: log-Gaussian Cox point processes

We now investigate the performance of the aMALA on a high-dimensional target density by studying the log-Gaussian Cox point process example presented in Christensen et al. (2005). The target density of this example possesses strong correlations and provides an interesting challenge for MCMC samplers. The dataset contains the locations of 126 Scots pine saplings in a Finnish natural forest. These locations are first standardized to fit on the region $[0,1]^2$, which becomes the area of interest. A discretization of this square is then obtained using a 64×64 regular grid, and the random variables $\mathbf{Y} = (Y_{i,j}, i,j = 1,\ldots,64)$ represent the number of points in each grid cell (i,j). Based on this fine discretization, the dimension of the model becomes $N = 64^2 = 4096$ and most grid cells are empty. The random variables are assumed to be conditionally independent given a latent intensity process $\Lambda(\cdot) = \{\Lambda(i,j) | i,j = 1,\ldots,64\}$ and are Poisson distributed with means $m\Lambda(i,j)$, $i,j = 1,\ldots,64$, where $m = 1/4096$ is the area of each cell. The latent intensity process is assumed to take the form $m\Lambda(i,j) = m\exp\{X_{i,j}\}$, where the prior for $\mathbf{X} = (X_{i,j}, i,j = 1,\ldots,64)$ is a multivariate Gaussian with mean $\mathbb{E}[\mathbf{X}] = \mu\mathbf{1}$ and $N \times N$ covariance matrix $\mathbb{C}\text{ov}(\mathbf{X}) = \Sigma$ such that

$$\mathbb{C}\text{ov}(X_{i,j}, X_{i',j'}) = \sigma^2 \exp\left\{-\frac{\sqrt{(i-i')^2 + (j-j')^2}}{64\beta}\right\} .$$

As in Christensen et al. (2005), we use the estimated hyperparameters $\beta = 1/33$, $\sigma^2 = 1.91$, and $\mu = \log(126) - \sigma^2/2$.

Using Bayes, the posterior density of interest satisfies

$$\pi(\mathbf{x}|\mathbf{y}) \quad \propto \quad \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu\mathbf{1})^\top \Sigma^{-1}(\mathbf{x}-\mu\mathbf{1})\right\} \prod_{i,j=1}^{64} \exp\left\{x_{i,j}y_{i,j} - m\exp\{x_{i,j}\}\right\} ,$$

for $\mathbf{x} \in \mathbb{R}^N$. The derivative of the log-density with respect to the latent variables is $\nabla \log \pi(\mathbf{x}|\mathbf{y}) = \mathbf{y} - m\exp\{\mathbf{x}\} - \Sigma^{-1}(\mathbf{x} - \mu\mathbf{1})$. As in Girolami and Calderhead (2011), we let the preconditioning matrix $\mathcal{C}_N = M^{-1}$ with $M = -\mathbb{E}[\nabla^2 \log \pi(\mathbf{x}|\mathbf{y})] = \Lambda + \Sigma^{-1}$, where Λ is an $N \times N$ diagonal matrix whose diagonal elements are obtained from the expectation of the exponential of normal random variables as $m\exp\{\mu + (\Sigma)_{ii}\}$, $i = 1,\ldots,4096$.

Given the pines dataset and fixed hyperparameters, we generate values from $\pi(\mathbf{x}|\mathbf{y})$ using a preconditioned aMALA with various combinations of initial value for $\mathbf{x}$ and tuning parameters γ, δ. We outline the fact that contrarily to the MALA of Christensen et al. (2005), it is not necessary to reparameterize the target density in order to benefit from an efficient exploration of the state space; we may thus use the density detailed above as is.

The initial values used in the simulations are $X_{i,j} = \mu$ for $i,j = 1,\ldots,64$, as well as a starting value near the posterior mode obtained by solving the equation $y_{i,j} - \exp\{X_{i,j}\} - (X_{i,j} - \beta)/\sigma^2 = 0$ for $X_{i,j}$, $i,j = 1,\ldots,64$; these values were studied in Christensen et al.

(2005). We also consider intermediate values (the vectors $5.657691 \cdot \mathbf{1}$ and $7.122988 \cdot \mathbf{1}$, which are the two largest values solving the previous equation), as well as initial values that are farther in the tail of the distribution, such as 64^2-dimensional vectors of 0's and 10's.

For each of these initial values, we implement MALA samplers ($\gamma = 1$) with tunings $\delta = 1/N^{1/2}$ and $\delta = 1.36/N^{1/3}$; the former is optimal in transience for a standard normal target started at the origin, while the latter is optimal in the stationary phase for a standard normal target. We also run aMALA with $\delta = \ell_1^2/N^{1/3}$, $\gamma = 1 + \ell_2^2/N^{1/3}$, and $\ell_2^2 = \ell_1^2/2$, where we investigate different values of ℓ_1^2. We study $\ell_1^2 = (2/3)^{(1/3)}$, which has been selected as near-optimal in transience according to Theorem 1. We also study a selection of ℓ_1^2 values ranging from 2 to 2.9, and finally implement aMALA with $\delta = 1.03/N^{1/5}$ and $\gamma = 1 + 0.515/N^{1/5}$, which has been found to be the optimal tuning in the stationary phase in §3.2.2. A summary of our findings is presented in Tables 1 and 2; for each algorithm implemented, we measure efficiency using the average squared jumping distance and we report the acceptance rate, as well as the initial value used.

				$\mathbf{X}_0 = \mu\mathbf{1}$		$\mathbf{X}_0$ near post. mode		$\mathbf{X}_0 = 0\mathbf{1}$	
Algorithm	ζ	ℓ_1^2	ℓ_2^2	ASJD	Acc. rate	ASJD	Acc. rate	ASJD	Acc. rate
MALA	1/2	1.00	0.00	154.15	0.967	154.16	0.967	154.23	0.967
MALA	1/3	1.36	0.00	0.00	0.000	0.00	0.000	0.00	0.000
aMALA	1/3	$(2/3)^{1/3}$	$(1/12)^{1/3}$	541.51	0.956	537.72	0.949	539.49	0.952
aMALA	1/3	2.00	1.00	1107.33	0.825	1104.69	0.823	0.00	0.000
aMALA	1/3	2.50	1.25	1264.02	0.742	1256.85	0.738	0.00	0.000
aMALA	1/3	2.62	1.31	1285.58	0.716	1274.07	0.707	0.00	0.000
aMALA	1/3	2.74	1.37	1315.83	0.697	1316.41	0.697	0.00	0.000
aMALA	1/3	2.90	1.45	1271.91	0.635	1287.19	0.643	0.00	0.000
aMALA	1/5	1.03	0.515	1178.64	0.543	1193.21	0.549	0.00	0.000

Table 1: 10,000 iterations of MALA and aMALA with the following initial values: μ, near posterior mode, and 0

The two samplers that are tuned so as to be optimal in transience (1st and 3rd line in each table) succeed in leaving their initial state and exploring the space in all cases, regardless of $\mathbf{X}_0$. As expected, the acceptance rates of these algorithms are quite large (97% for MALA and 95% for aMALA), which indicates that a tuning adjustment should probably be performed once the process leaves the transient phase. In all cases, aMALA offers a performance that is about 3.5 times as good as that of MALA (improvement ranging between 348% and 352% in terms of ASJD). This is explained by the fact that aMALA explores its state space in $\mathcal{O}(N^{1/3})$ iterations, compared to $\mathcal{O}(N^{1/2})$ for MALA.

As the process is started further out in the tails, it becomes increasingly difficult for other versions of these algorithms (with different tunings ℓ_1^2) to leave the initial state. In

				$\mathbf{X}_0 = 5.657691\mathbf{1}$		$\mathbf{X}_0 = 7.122988\mathbf{1}$		$\mathbf{X}_0 = 10\mathbf{1}$	
Algorithm	ζ	ℓ_1^2	ℓ_2^2	ASJD	Acc. rate	ASJD	Acc. rate	ASJD	Acc. rate
MALA	1/2	1.00	0.00	153.96	0.966	154.02	0.967	154.59	0.967
MALA	1/3	1.36	0.00	0.00	0.000	508.48	0.570	0.00	0.000
aMALA	1/3	$(2/3)^{1/3}$	$(1/12)^{1/3}$	539.66	0.952	536.56	0.946	544.81	0.955
aMALA	1/3	2.00	1.00	1075.85	0.802	1108.08	0.825	0.00	0.000
aMALA	1/3	2.50	1.25	1076.49	0.632	1106.60	0.649	0.00	0.000
aMALA	1/3	2.62	1.31	0.18	1e-04	1032.91	0.574	0.00	0.000
aMALA	1/3	2.74	1.37	0.00	0.000	345.28	0.183	0.00	0.000
aMALA	1/3	2.90	1.45	0.00	0.000	190.85	0.095	0.00	0.000
aMALA	1/5	1.03	0.515	0.00	0.000	0.41	1e-04	0.00	0.000

Table 2: 10,000 iterations of MALA and aMALA with the following initial values: 5.657691, 7.122988, and 10

spite of this, we notice that aMALA does much better than MALA, which almost never leaves its initial value when tuned according to its stationary settings. We emphasize the fact that the target distribution studied does not satisfy the regularity conditions expounded in §2.4 (the covariance matrix used in aMALA is only an approximation, among other things). Consequently, the asymptotically optimal tuning $\ell_1^2 = 1.03$ does not necessarily corresponds to a 70% acceptance rate. According to the numerical results obtained, it would appear that adjusting the acceptance rate is a tuning approach more robust than adjusting ℓ_1^2. Indeed, the ASJD is near its maximum when the acceptance rate is tuned so as to approach 70%, so this is probably what users should aim for. In fact, when tuning aMALA so that it accepts 70% of candidates, we realize that it only fails to leave the initial value when the latter is extreme (0's or 10's). We also note that an aMALA that is optimally tuned for its stationary phase is more than twice as efficient as MALA in terms of ASJD (when the initial state is 7.122988, which is the only occurence of MALA exploring its state space with its $\mathcal{O}(N^{-1/3})$ optimal tuning). Specifically, a run of aMALA with $\ell_1^2 = 2.4$ leads to an ASJD of 1211.71 and an acceptance rate of 74.3%, which represents an improvement of 240% over the MALA tuned to accept 57% of candidates (extra run not recorded in the tables).

6 Discussion

This paper presents asymptotic results about the new MALA with annealed proposal, both in and out of stationarity. As was done in the literature for MALA, we consider a preconditioned version of aMALA and study its computational cost in transience and in stationarity. Through the various developments, it becomes obvious that aMALA performs best when its tuning parameters satisfy a specific relationship: $\delta = \ell_1^2/N^{\zeta_1}$ and $\gamma =$

$1 + \ell_2^2/N^{\zeta_2}$, with $\ell_2^2 = \ell_1^2/2$ and $\zeta_1 = \zeta_2$. This implies that γ is entirely determined by our choices of ℓ_1^2 and ζ_1 for δ, which ultimately leads to a sampler as easy to tune as MALA. We find that aMALA's computational cost out of the stationary phase is $\mathcal{O}(N^{1/3})$, a significant improvement over MALA's $\mathcal{O}(N^{1/2})$ cost. Although there is no value ℓ_1^2 that optimizes the speed of convergence in transience, the choice $\ell_1^2 = (2/3)^{1/3}$ seems reliable and close to optimality for a broad range of initial values.

In its stationary phase, the preconditioned aMALA is found to have a computational cost of $\mathcal{O}(N^{1/5})$, provided that $\ell_2^2 = \ell_1^2/2$ holds. In that case, one can simply tune ℓ_1^2 so as to approach the asymptotically optimal acceptance rate of 70.4%. If a user were to select another value of ℓ_2^2, violating the above relation, then the computational cost would become $\mathcal{O}(N^{1/3})$ and the asymptotically optimal acceptance rate would then be smaller than 70% (but an exact optimal value would be difficult to find). Nonetheless, simulation studies show that aMALA offers better performances than MALA in that case, as long as ℓ_1^2 is not too small (in practice, we can easily remediate to this potential problem by using a few preliminary runs with different ℓ_1^2 values).

Finally, when omitting the preconditioning matrix in aMALA, we find that the parameter ℓ_2^2 should be chosen so as to satisfy (13). In that case, the computational cost is $\mathcal{O}(N^{1/3})$ and ℓ_1^2 should simply be tuned so as to target a 57.4% acceptance rate in the stationary phase, as is the case for MALA. The new sampler however leads to better efficiency measures (i.e. a larger speed measure) due to the presence of γ in the proposal mean. In all the situations considered, the theoretical and practical results obtained in this paper are in agreement and unequivocally show the advantages of implementing aMALA over MALA.

Statements and Declarations

Competing interests: none
Data: The dataset `finpines` is publicly available in `spatstat.data` on `R`
Code: `R` code will be made available, TBA.
Supplementary material: Appendices A to D, containing the proof of Theorem 1

References

Boisvert-Beaudry, G. and M. Bédard (2022). MALA with annealed proposals: A generalization of locally and globally balanced proposal distributions. *Statistics and Computing 32*(1), 5.

Christensen, O. F., G. O. Roberts, and J. S. Rosenthal (2005). Scaling limits for the transient phase of local Metropolis–Hastings algorithms. *Journal of the Royal Statistical Society Series B: Statistical Methodology 67*(2), 253–268.

Girolami, M. and B. Calderhead (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(2), 123–214.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*(1), 97–109.

Kuntz, J., M. Ottobre, and A. M. Stuart (2018). Non-stationary phase of the MALA algorithm. *Stochastics and Partial Differential Equations: Analysis and Computations 6*, 446–499.

Kuntz, J., M. Ottobre, and A. M. Stuart (2019). Diffusion limit for the random walk Metropolis algorithm out of stationarity. *Annales de l'Institut Henri Poincaré - Probabilités et Statistique 55*(3), 1599–1648.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21*(6), 1087–1092.

Roberts, G. O., A. Gelman, W. R. Gilks, et al. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability 7*(1), 110–120.

Roberts, G. O. and J. S. Rosenthal (1998). Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60*(1), 255–268.

Zanella, G. (2020). Informed proposals for local MCMC in discrete spaces. *Journal of the American Statistical Association 115*(530), 852–865.