Noname manuscript No.
(will be inserted by the editor)

MALA with annealed proposals

A generalization of locally- and globally-balanced proposal distributions

Gabriel Boisvert-Beaudry · Mylène Bédard

Received: date / Accepted: date

Abstract We introduce a generalized version of the Metropolis-adjusted Langevin algorithm (MALA). The informed proposal distribution of this new sampler features two tuning parameters: the usual step size parameter σ^2 and an interpolation parameter γ that may be adjusted to accommodate the dimension of the target distribution. We theoretically study the efficiency of the sampler by making use of the local- and global-balance concepts introduced in Zanella (2020) and provide efficient tuning guidelines that work well with a variety of target distributions. Although the usual MALA ($\gamma = 1$) is shown to be optimal for infinite-dimensional targets, in practice, the generalized MALA ($1 < \gamma \leq 2$) remains the most appealing option, even in high-dimensional contexts. Simulation studies and numerical experiments are presented to illustrate our findings. We apply the new sampler to a Bayesian logistic regression context and show that its efficiency compares favourably to competing algorithms.

Keywords Bayesian logistic regression · Informed proposal distribution · Markov chain Monte Carlo · Metropolis-adjusted Langevin algorithm · Reversibility

Mathematics Subject Classification (2010) 60J22 · 62F15

1 Introduction

Statistical models to study real-world phenomena have been increasing both in terms of complexity and dimensionality. Such models generally produce densities that cannot be treated analytically; MCMC methods have thus become a device of choice to obtain samples from these complicated probability distributions.

The Metropolis–Hastings sampler (Metropolis et al., 1953; Hastings, 1970) is at the core of the MCMC toolbox. The idea is to build a Markov process with invariant distribution Π on a state space $\mathcal{S}$ by proposing candidates to be included in the process according to some acceptance probability. Let an initial value X_0 for the process be drawn from an arbitrary distribution μ and let π be the d-dimensional target density arising from Π with respect to Lebesgue measure. Then, at iteration $t+1$, the Metropolis–Hastings (MH) sampler generates a candidate $Y_{t+1} = y$ from a proposal distribution $Q(X_t, y)$ with density $q(X_t, y)$. This candidate is accepted as the next state X_{t+1} of the Markov process with probability $\alpha(X_t, y) = \min\{1, \frac{\pi(y)q(y,X_t)}{\pi(X_t)q(X_t,y)}\}$, otherwise we set $X_{t+1} = X_t$ and the process remains at the current state for another time interval.

The role of the acceptance probability $\alpha(x, y)$ is one of correction: it makes sure that accepted candidates, which are generated using the proposal distribution, can be considered as coming from the target distribution Π. This probability is chosen so as to make the Markov process time-reversible with respect to Π, that is to satisfy the detailed balance condition

$$\pi(x)q(x, y)\alpha(x, y) = \pi(y)q(y, x)\alpha(y, x) \ ,$$

for all $x, y \in \mathcal{S}$. For discrete-space Markov chains, this intuitively means that the probability of moving from x at time t to y at time $t+1$ is equal to the probability

The authors are grateful to the two anonymous reviewers whose suggestions and comments led to an improved manuscript. They also acknowledge the support of the *Natural Sciences and Engineering Research Council of Canada.*

Université de Montréal,
(Département de mathématiques et de statistique),
Montréal, (QC), Canada
E-mail: mylene.bedard@umontreal.ca

of the reverse move. Conveniently, if a Markov chain is reversible with respect to $\varPi$, it implies that $\varPi$ is stationary for the chain.

A generic choice of proposal distribution is to draw candidates from a $\mathcal{N}(X_t, \sigma_d^2 I_d)$, where I_d is the $d \times d$ identity matrix and $\sigma_d > 0$ is a scalar for tuning; this yields a random walk version of the MH sampler (RWMH). This simple distribution uses no $\varPi$-related information to guide the exploration of the state space and, in that sense, is thus blinded. For the Markov process to rapidly explore its state space $\mathcal{S}$, careful tuning of the Q parameters is required. Optimal tuning is however not a cure-all for basic proposal designs; even carefully tuned RWMH samplers sometimes fail, in practice, to appropriately explore the distribution of interest. This is often the case with distributions formed of highly correlated components, for instance.

Seeing as the previous proposal distribution is blinded, we could instead opt for an informed proposal distribution, one that uses the target $\varPi$ to guide the exploration of the state space. Such distributions propose, on average, better candidates; the acceptance probability thus has less correcting to do. In fact, the more informed is the proposal distribution, the fewer corrections are required, and the higher is $\alpha(x, y)$ under optimal tuning. Taking this to the extreme, if we were able to sample directly from $\varPi$, then we could choose $q(x, y) = \pi(y)$ and no correction would be needed as $\alpha(x, y) = 1$. One example of an informed proposal distribution is arising in the Metropolis-adjusted Langevin algorithm (MALA), where the gradient of the target's log-density directs the search of candidates towards regions of high probability. The MALA is a Metropolis-Hastings algorithm with proposal distribution

$$Y_{t+1} \sim \mathcal{N}\left(X_t + \frac{\sigma_d^2}{2} \nabla \log\left\{\pi(X_t)\right\}, \sigma_d^2 I_d\right) ,$$

with $\sigma_d > 0$ for tuning.

Under certain regularity conditions on the target density, it has been proven that optimally tuned versions of the RWMH accept 23.4% of candidates and explore their state space in $\mathcal{O}(d)$ iterations (Roberts et al., 1997); similar results for the MALA tell that the optimal acceptance rate is 57.4% and exploration of $\mathcal{S}$ is achieved in $\mathcal{O}(d^{1/3})$ iterations (Roberts and Rosenthal, 1998). These optimal scaling results are valid asymptotically (as the target dimension $d \uparrow \infty$) and illustrate that there are significant efficiency gains available from using informed proposal distributions by opposition to blinded ones. Such benefits are naturally not free, and may come at prices that are more or less expensive in terms of computational effort.

The ultimate goal is then to find a proposal density that requires the less amount of correction possible from the acceptance function, i.e. a proposal similar to π. Naturally, the only way to omit the acceptance probability in the detailed balance condition would be to sample directly from π. This is generally not a viable option as this density is complicated, from where the need to turn to samplers such as RWMH and MALA. Zanella (2020) however expounds a really nice theory about locally- and globally-balanced proposal distributions. These concepts may be seen as weaker forms of reversibility, where the detailed balance condition holds in limiting cases only (i.e. when $\sigma_d \downarrow 0$ or $\sigma_d \uparrow \infty$). He also introduces a general class of biased proposal distributions and provides conditions under which it satisfies the local and global balance conditions.

Although motivated in a discrete-space context, the theory is also applicable in the continuous case; the framework is presented in §2. Building on these foundations, we thus combine locally- and globally-balanced proposal kernels in order to propose, in §3, an annealed version of the usual MALA that features an extra tuning parameter. Theoretical results about the efficiency of this sampler are introduced in §3.1 and §3.2, along with some tuning guidelines. Significant efficiency gains come at no expense (with reference to MALA), as illustrated in the simulation studies of §3.3 and §3.4, as well as in the real data examples of §4.

2 Developing informed proposal kernels

Let $Q_\sigma(x, \cdot)$ be a symmetrical proposal distribution centered at x, with scaling parameter σ. Now, let $g : \mathcal{S} \times \mathcal{S} \to [0, \infty)$ be a bounded, continuous function. Zanella (2020) proposes to use the function g to bias the blinded kernel Q_σ; the resulting biased proposal distribution thus satisfies

$$Q_{g,\sigma}(x, \mathrm{d}y) = \frac{g(x, y) Q_\sigma(x, \mathrm{d}y)}{Z_{g,\sigma}(x)} , \quad \forall y \in \mathcal{S} , \tag{1}$$

where $Z_{g,\sigma}(x) = \int_{\mathcal{S}} g(x, z) Q_\sigma(x, \mathrm{d}z)$ is a normalizing constant. When there is no confusion about our choice of g, we use the lighter notation $Z_\sigma(x)$; similarly, when σ is fixed, we use $Z_g(x)$.

Interesting choices for g contain some information about the target density π. For instance, one could "transfer" the acceptance function of the RWMH into the proposal kernel by setting $g(x, y) = \pi(y)/\pi(x)$; in that case, $Q_{g,\sigma}(x, \mathrm{d}y) \propto \pi(y) Q_\sigma(x, \mathrm{d}y)$. Naturally, if π is complicated, it might not be easy to obtain candidates from this distribution, but we worry later as to how this could be achieved. If we instead select $g(x, y) = 1$, we are back to the blinded proposal

$Q_{g,\sigma}(x, \mathrm{d}y) = Q_\sigma(x, \mathrm{d}y)$. Of interest in thus how to choose the biasing function g. In particular, could we find a function g that almost eliminates the need for an accept/reject step?

2.1 Local and global balances

To completely avoid the accept/reject step in the MH sampler while simultaneously satisfying the detailed balance condition $\pi(x)q(x, y) = \pi(y)q(y, x)$ for all $x, y \in \mathcal{S}$, we need to sample directly from Π, which seems infeasible. Since it is impossible to find an alternative proposal kernel that satisfies the detailed balanced condition by itself, without any help from the acceptance function, Zanella (2020) introduces weaker concepts of balance.

Local/global balances Let $\{Q_\sigma\}_{\sigma>0}$ be a family of transition kernels. We say that $\{Q_\sigma\}_{\sigma>0}$ is locally-balanced with respect to a target Π if, for every Q_σ, the Markov process is reversible with respect to a distribution Π_σ such that $\Pi_\sigma \to \Pi$ weakly as $\sigma \downarrow 0$. Similarly, we say that $\{Q_\sigma\}_{\sigma>0}$ is globally-balanced with respect to Π if instead $\Pi_\sigma \to \Pi$ weakly as $\sigma \uparrow \infty$.

These concepts state that reversibility is attained in limiting cases only, where the algorithm then samples directly from the target distribution, without a need to submit candidates to an accept/reject step.

As it turns out, locally-balanced schemes, which become reversible as $\sigma \downarrow 0$, are appropriate in high-dimensional contexts: as d increases, smaller σ values are required to avoid facing an acceptance rate that converges to 0. Globally-balanced schemes, for their part, are more of a theoretical concept as one never really requires infinitely large tuning parameters. Nonetheless, they should be more appropriate in very small-dimensional contexts, that is when σ is as large as it may be.

Specifically, let σ_1^2 be the optimal scaling parameter of the RWMH (or MALA) for sampling from some one-dimensional target; as higher-dimensional versions of this target are studied, the optimal σ_d^2 decreases from σ_1^2 to eventually $\sigma_\infty^2 = 0$. This drop may happen more or less rapidly, depending for instance on the correlation among target components as d increases. In any case, σ_d^2 is a decreasing function of d and unidimensional targets are the closest we can naturally get to the globally-balanced case. In particular, Roberts et al. (1997) and Roberts and Rosenthal (1998) show under certain regularity conditions that scaling parameters should be of the form $\sigma_d^2 = \ell^2/d$ and $\sigma_d^2 = \ell^2/d^{1/3}$ with $\ell > 0$ for RWMH and MALA, respectively. Following this line of reasoning, relying on the right regime (local vs. global) in a specific d-dimensional context should lead to better, more frequently accepted candidates.

2.2 Selecting the balancing function

There remains the problem of actually selecting the right function g. This function can be quite general, but we mimic Zanella (2020) and focus on functions of x, y through the ratio $\pi(y)/\pi(x)$. To avoid integrability issues, assume $\tilde{g} : [0, \infty) \to [0, \infty)$ to be continuous and bounded by some linear function, meaning that $\tilde{g}(t) \leq a + bt$ for some $a, b > 0$ and all $t \geq 0$. Under these assumptions, Zanella (2020) derives conditions on g that lead to locally-balanced algorithms. In his Theorem 1, he shows the local balance condition to be satisfied if, and only if, $g(x, y) = \tilde{g}(\pi(y)/\pi(x))$, with

$$\tilde{g}\left(\frac{\pi(y)}{\pi(x)}\right) = \frac{\pi(y)}{\pi(x)}\tilde{g}\left(\frac{\pi(x)}{\pi(y)}\right) , \quad \forall x, y \in \mathcal{S} . \tag{2}$$

This nicely illustrates the duality between the acceptance function and the proposal kernel.

Although Zanella (2020) only briefly addresses the concept of globally-balanced algorithms, it is not too difficult to show that the global balance condition is satisfied if, and only if,

$$g(x, y) \propto \pi(y) , \quad \forall x, y \in \mathcal{S} ;$$

here, the proportionality only concerns y as the biasing function is ultimately involved in a proposal distribution for y. The proof of this result may be found in §A.1 of Appendix A. Intuitively, as $\sigma \uparrow \infty$, the blinded kernel Q_σ in (1) becomes flat and so to attain reversibility, we must sample directly from Π. The proposal distribution thus satisfies $Q_{g,\sigma}(x, \mathrm{d}y) \propto g(x, y)Q_\sigma(x, \mathrm{d}y) \propto \Pi(\mathrm{d}y)$ and the acceptance rate becomes equal to 1 as candidates are automatically included in the sampler.

Although finding practically-implementable globally-balanced kernels is generally difficult, in some idealised sense, such kernels still are desirable from a theoretical viewpoint. In fact, we can show that the expected squared jumping distance of the Markov process, which measures the efficiency of the sampler in exploring its state space, remains positive in a globally-balanced context despite an infinitely large tuning parameter σ.

Proposition 1 *Let π be a bounded target density such that $\mathbb{E}_\pi[\|X\|^2] < \infty$, where $\|\cdot\|$ denotes the Euclidean norm. Suppose that a Metropolis-Hastings algorithm with a globally-balanced proposal distribution is used to sample from this target. Let the blinded portion of the proposal density, q_σ, be such that*

$$\pi(y) > 0 \;\Rightarrow\; q_\sigma(x, y) > 0 , \quad \forall x \in \mathcal{S} .$$

Then,

$$\lim_{\sigma\to\infty}\mathbb{E}[\|X_{t+1}-X_t\|^2] =$$

$$\lim_{\sigma\to\infty}\iint \|y-x\|^2 q_{g,\sigma}(x,y)\alpha(x,y)\mathrm{d}y\mathrm{d}x > 0 \ ,$$

where $q_{g,\sigma}$ is the density associated to $Q_{g,\sigma}$.

Proof The proof of this result may be found in §A.2 of Appendix A or in §3.2 of Boisvert-Beaudry (2019).

While the conclusion of this result is very much intuitive, we outline the fact that neither RWMH nor MALA achieves as much when $\sigma \uparrow \infty$. In both cases, the acceptance rate simply converges to 0 and so does the expected squared jumping distance, leaving us with a dead process. In low-dimensional settings, where we tend to use relatively large σ, globally-balanced proposal kernels are thus expected to be more appropriate, or closer to optimality, than locally-balanced ones.

2.3 Asymptotic efficiency of locally-balanced proposals

Having characterized globally- and locally-balanced proposal kernels, it would be useful to understand to which extent locally-balanced distributions become more efficient than other biased kernels as $d \uparrow \infty$. Although the theoretical results in this section are introduced in a discrete-space framework, which is more accessible in terms of exposition, we note that they also hold in continuous-space settings.

By focusing on a discrete-space framework, Zanella (2020) introduces some theoretical results for measuring the efficiency of MCMC samplers and comparing their convergence properties. These results are based on the concepts of spectral gap and asymptotic variance of an arbitrary function $h : \mathcal{S} \to \mathbb{R}$ (with $\mathcal{S}$ discrete). The asymptotic variance is defined as

$$\mathbb{V}ar_\pi(h,P) = \lim_{N\uparrow\infty}\frac{1}{N}\mathbb{V}ar\left(\sum_{t=1}^{N} h(X_t)\right) \ ,$$

where $\{X_t; t \geq 0\}$ is a Markov chain with transition kernel P started in the stationary distribution Π. The smaller is the asymptotic variance of some function h, the less correlation there is among MCMC samples and the more efficient is the sampler in estimating $\mathbb{E}_\pi[h]$, the expectation of $h(X)$ when $X \sim \Pi$. The spectral gap $Gap(P) = 1 - \lambda_2 \geq 0$, where λ_2 is the second largest eigenvalue of P, may also be used to compare the convergence of two samplers; the further from 0 is the value of the spectral gap, the fastest is the convergence of the Markov chain.

In his Theorem 2, Zanella (2020) shows that if P_1 and P_2 are π-reversible Markov transition kernels on $\mathcal{S}$ with $P_1(x,y) \geq cP_2(x,y)$ for all $x \neq y$ and a fixed $c > 0$, then $Gap(P_1) \geq c\ Gap(P_2)$ and

$$\mathbb{V}ar_\pi(h,P_1) \leq \frac{1}{c}\mathbb{V}ar_\pi(h,P_2) + \frac{1-c}{c}\mathbb{V}ar_\pi(h) \qquad (3)$$

for all $h : \mathcal{S} \to \mathbb{R}$, where $\mathbb{V}ar_\pi(h)$ is the variance of $h(X)$ when X has density π.

This result says that when the inequality of the matrices holds, then the transition kernel P_1 is c times more efficient than P_2 in terms of asymptotic variance and spectral gap (the term $\mathbb{V}ar_\pi(h)$ in (3) is usually much smaller than the other term in that equation). In particular, the case $c = 1$ is known as Peskun ordering; we refer the reader to Peskun (1973) and Tierney (1998) for more details about Peskun orderings on discrete and continuous state spaces, respectively.

Now, let $c_g = \sup_{(x,y)\in S}\{Z_g(y)/Z_g(x)\}$, where $S = \{(x,y) \in \mathcal{S}\times\mathcal{S} : \pi(x)q(x,y) > 0\}$, $Z_g(x)$ is the normalizing constant in (1), and $q(x,y)$ is the density of $Q_\sigma(x,y)$ for a fixed σ. The term c_g is the largest possible ratio of normalizing constants over any two potential consecutive states x and y; by construction, this ratio is ≥ 1. In its Theorem 3, Zanella (2020) shows how to asymptotically improve the efficiency of a MH sampler that uses a proposal distribution as in (1) with a biaising function $\tilde{g}(\pi(y)/\pi(x))$. In particular, he defines $\hat{g}(t) = \min\{\tilde{g}(t), t\tilde{g}(1/t)\}$ and demonstrates that if $P_{\tilde{g}}$ and $P_{\hat{g}}$ are the MH kernels obtained from the biased proposal kernels $Q_{\tilde{g}}$ and $Q_{\hat{g}}$ respectively, then

$$P_{\hat{g}}(x,y) \geq \frac{1}{c_{\tilde{g}}c_{\hat{g}}}P_{\tilde{g}}(x,y) \ , \qquad \forall x \neq y \ .$$

It turns out that $\hat{g}$ satisfies $\hat{g}(t) = t\hat{g}(1/t)$; therefore, for any biasing function $\tilde{g} : [0,\infty) \to [0,\infty)$, there is a corresponding $\hat{g}$ that leads to a locally-balanced kernel $Q_{\hat{g}}$. Since $c_{\tilde{g}}, c_{\hat{g}} \geq 1$, Theorems 2 and 3 of Zanella (2020) *do not* imply that $P_{\hat{g}}$ is better than $P_{\tilde{g}}$. In many cases however, the term $(c_{\tilde{g}}c_{\hat{g}})^{-1}$ converges to 1 as $d \uparrow \infty$; when this is true, this means that the locally-balanced kernel $P_{\hat{g}}$ is asymptotically optimal in terms of Peskun ordering. Therefore, as $d \uparrow \infty$, $P_{\tilde{g}}$ cannot significantly improve over $P_{\hat{g}}$ in terms of asymptotic variance and convergence (spectral gap).

Accordingly, locally-balanced proposals produce MH algorithms that are asymptotically optimal within the class of proposal distributions (1) with biasing function $\tilde{g}(\pi(y)/\pi(x))$. We note that $\hat{g}$ cannot be used to asymptotically improve on $\tilde{g}$ if the latter already is locally balanced. Indeed, in that case, $t\hat{g}(1/t) = \hat{g}(t) = \tilde{g}(t) = t\tilde{g}(1/t)$ and $P_{\hat{g}} = P_{\tilde{g}}$. In fact, there is generally no choice of g that Peskun-dominates the others when restricted to biaising functions satisfying $\tilde{g}(t) = t\tilde{g}(1/t)$.

Now, let

$$b_{\tilde{g}} = \sup_{(x,y)\in S}\tilde{g}(\tfrac{\pi(y)}{\pi(x)})/\{(\tfrac{\pi(y)}{\pi(x)})\tilde{g}(\tfrac{\pi(x)}{\pi(y)})\} \ ,$$

with S as before. The term $b_{\tilde{g}} \geq 1$ measures how unbalanced' is the biasing function $\tilde{g}$: the larger is $b_{\tilde{g}}$, the more unbalanced is $\tilde{g}$ with respect to $\tilde{g}(t) = t\tilde{g}(1/t)$; when $b_{\tilde{g}} = 1$, then the previous relation is satisfied. In his Theorem 4, Zanella (2020) shows that $P_{\tilde{g}}(x,y) \geq P_{\hat{g}}(x,y)/\{c_{\tilde{g}}c_{\hat{g}}b_{\tilde{g}}\}$ for all $x \neq y$. This means that the less balanced is a function $\tilde{g}$, the more improvement we can reasonably expect from relying on a locally-balanced kernel as $d \uparrow \infty$. Furthermore, for a rough target, the expected improvement from using a locally-balanced proposal is greater than for a smooth target.

In finite-dimensional settings, we expect conclusions to be different from those just outlined. Theorems 2 and 3 of Zanella (2020) claim that $\hat{g}$ is $\{c_{\tilde{g}}c_{\hat{g}}\}^{-1}$ times more efficient than some biased function $\tilde{g}$. Since $c_{\tilde{g}}, c_{\hat{g}} \geq 1$, we believe we can find a biasing function $\tilde{g}$ that offers better performances than $\hat{g}$ in small to moderate dimensions, with $\hat{g}$ eventually becoming more efficient than $\tilde{g}$ when d is large enough (as $c_{\tilde{g}}c_{\hat{g}}$ approaches 1). From Theorem 4, the extent of the improvement of $\tilde{g}$ over $\hat{g}$ will not only depend on the choice of $\tilde{g}$, but also on the roughness of the target studied.

2.4 Approximating the balancing function

As g typically is a function of the target density π, sampling from the above-mentioned biased proposal kernels does not seem accessible. We need to resort to approximation schemes to obtain manageable distributions. Hereafter, we suppose the uninformed kernel Q_σ to be the $\mathcal{N}(x, \sigma^2 I_d)$ of the RWMH.

In the locally-balanced context, it is relatively easy to find biasing functions $g(x,y) = \tilde{g}(\pi(y)/\pi(x))$ that satisfy (2). Using a first-order Taylor approximation to reexpress the function g in $Q_{g,\sigma}$, it turns out that at least two possibilities for the balancing function lead to the usual MALA sampler, that is

$$g(x,y) = \left\{\frac{\pi(y)}{\pi(x)}\right\}^{1/2} \quad \text{and} \quad g(x,y) = \frac{\pi(y)/\pi(x)}{1+\pi(y)/\pi(x)} \ .$$

Details of the calculations can be found in Appendix A.3 (the square root balancing function is also discussed in §5 of Zanella, 2020). This satisfyingly confirms the efficiency of MALA in high-dimensional contexts. Naturally, as local balance is only attained in the limit when $\sigma \downarrow 0$, and since MALA is an approximation to genuine locally-balanced kernels, then candidates are still submitted to an accept/reject step before being included in the process.

In the globally-balanced context, a biasing function $g(x,y) \propto \pi(y)$ is required ($\propto$ is with respect to y only). Since the case $\sigma \uparrow \infty$ is purely theoretical and virtually never arises in practice, we do not elaborate on this but still inspire ourselves from this balancing function to propose a general form for g. Let us think of $g(x,y)$ as a biasing function that minimizes the amount of correction coming from the acceptance function. For infinitely small – or large – σ, we identified functions $g(x,y)$ that eliminate the need for such corrections through the usual accept/reject step. Now for intermediate cases where $0 < \sigma < \infty$, such a perfect biasing function $g(x,y)$ might or might not be available; nevertheless, it still makes sense to look for a function that minimizes – or at least diminishes significantly – the amount of correction required from the acceptance function. Seeing as $g(x,y) \propto \{\pi(y)\}^{1/2}$ for $\sigma \downarrow 0$ and $g(x,y) \propto \{\pi(y)\}^1$ for $\sigma \uparrow \infty$, we introduce the generalized biasing function $g(x,y) \propto \{\pi(y)\}^{\gamma/2}$ with $\gamma \in [1,2]$. Following the above theory, we expect this kernel to be more appropriate than locally- and globally-balanced kernels when $0 < \sigma < \infty$, which reasonably corresponds to any finite-dimensional target.

The generalized g leads to the proposal density

$$\begin{aligned} &q_{g,\sigma}(x,y)\mathrm{d}y \\ &\propto g(x,y)q_\sigma(x,y)\mathrm{d}y \\ &\propto \{\pi(y)\}^{\gamma/2} \exp\left\{-\frac{1}{2\sigma^2}(y-x)^\top(y-x)\right\}\mathrm{d}y \ . \end{aligned}$$

Using a first-order Taylor approximation around x to develop $g(x,y) \propto \{\pi(y)\}^{\gamma/2}$, we find

$$\{\pi(y)\}^{\gamma/2} \approx \exp\left\{\frac{\gamma}{2}\log\{\pi(x)\} + \frac{\gamma}{2}\nabla\log\{\pi(x)\}(y-x)\right\} \ ;$$

the approximated biased proposal density is then

$$\begin{aligned} &q_{g,\sigma}(x,y)\mathrm{d}y \\ &\propto \exp\left\{\frac{\gamma}{2}\nabla\log\{\pi(x)\}(y-x)\right\} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma^2}(y-x)^\top(y-x)\right\}\mathrm{d}y \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\left\|y-x-\frac{\gamma\sigma^2}{2}\nabla\log\{\pi(x)\}\right\|^2\right\}\mathrm{d}y \ , \end{aligned}$$

leading to the proposal distribution

$$Y_{t+1} \sim \mathcal{N}\left(X_t + \frac{\gamma\sigma^2}{2}\nabla\log\{\pi(X_t)\}, \sigma^2 I_d\right) \ . \tag{4}$$

This proposal kernel is very similar to that of MALA, the only difference consisting in the extra parameter γ in front of the gradient term. Although this might seem like a minor change, we will realize shortly that this modification often leads to significant efficiency gains with respect to MALA. Naturally, candidates generated using the above kernel still need to go through an accept/reject step to preserve the sampler's reversibility.

3 The annealed MALA

The generalized kernel introduced above, which is expected to better accommodate finite-dimensional contexts than the usual MALA, leads to a sampler hereafter labeled as 'MALA with annealed proposals', or simply 'annealed MALA'. In particular, let $\gamma \in [1, 2]$ and $\sigma > 0$; the full density associated to the proposal distribution in (4) is defined as

$$q_{\gamma,\sigma}(x, y) = (2\pi\sigma^2)^{-d/2} \times \exp\left\{ -\frac{1}{2\sigma^2} \left\| y - x - \frac{\gamma\sigma^2}{2} \nabla \log\{\pi(x)\} \right\|^2 \right\} , \tag{5}$$

for $y \in \mathbb{R}^d$ and $x \in \mathcal{S}$. We can think of γ as an interpolation parameter since it makes a connection between the local and global balancing functions of §2.4. Indeed, the biasing function $g(x, y) \propto \{\pi(y)\}^{\gamma/2}$ may find itself closer to the local equilibrium ($\gamma = 1$), the global one ($\gamma = 2$), or anywhere in between. We note that setting $\gamma = 0$ leads back to the traditional blinded RWMH sampler but, in the current context, it seems inappropriate to include cases outside of the local-global range of proposal kernels.

The interpolation parameter γ adds to the versatility of the proposal kernel. By imposing $\gamma = 1$ as in MALA, for instance, users automatically freeze the relationship between proposal variance and proposal mean. By subsequently tuning σ, they thus simultaneously set (1) the variability of candidates around the mean of the proposal kernel, and (2) the weight in front of the gradient term, itself part of the proposal mean. Therefore, as σ increases, the variability among candidates not only increases, but the latter also gravitate around a point that moves towards high probability regions in a predetermined fashion. There is however no guarantee that the coveted high-density regions are attained by the mean (we might require a more aggressive biasing of the gradient), or are not already far behind (which would call for a smaller factor in front of the gradient). Indeed, different targets might require distinct biasing strategies; this is where the interpolation parameter makes a difference.

According to the preceding local and global balances theory, the MALA kernel appears as overly conservative when it comes to the biasing of its proposal mean. This is unsurprising as the initial design of MALA originates from the discretization of a Langevin diffusion process which, broadly speaking, corresponds to the limiting behaviour of infinite-dimensional MH algorithms as $\sigma \downarrow 0$. As hinted by the global biasing function one should therefore assign more weight to the informed portion of the proposal kernel (here the gradient term). All else being equal, smaller-dimensional proposal distributions should therefore generate candidates that are more biased than higher-dimensional ones.

Naturally, where augmented versatility comes into play, a greater attention to tuning is also required. In addition to finding the optimal σ value, one also needs to select an appropriate value for the interpolation parameter γ. The new proposal distribution thus calls for some guidance with respect to the tuning of this extra parameter. Since γ is understood to increase with σ, we expect the optimal γ to be a decreasing function of the dimension. Before saying more about the tuning of this parameter, we first settle the matter of σ.

3.1 Tuning the proposal variance

As part of optimizing the performances of the annealed MALA, we study the tuning of its step size parameter σ. On the one hand, aggressive candidate steps (large σ_d^2) tend to always be rejected and lead to a process that is frozen, more often than not. On the other hand, conservative candidates (small σ_d^2) curb the process, which then necessitates too many iterations for travelling across the state space. We thus aim at striking a balance between sizeable steps and accepted candidates. For a given proposal distribution, the optimal form of the proposal variance can be expressed as $\sigma_d^2 = \ell^2/d^{\beta_0}$, with

$$\beta_0 = \min_{\beta_c \geq 0} \left\{ \beta_c : \lim_{d\to\infty} \mathbb{E}[\alpha(X, Y)] > 0, \quad \forall \beta \in [\beta_c, \infty) \right\} .$$

The scaling β_0 therefore leads to the largest possible σ_d^2 featuring a positive expected acceptance rate for all $d \geq 1$.

It has been mentioned previously that with a tuning parameter of the form $\sigma_d^2 = \ell^2/d^{1/3}$ ($\ell > 0$), MALA explores its space in $\mathcal{O}(d^{1/3})$ iterations. This turns out to be the largest proposal variance, in terms of d, that ensures a positive expected acceptance rate for all d for this sampler ($\beta_0 = 1/3$). This means that when $\beta_0 < 1/3$, then $\sigma_d^2 = \ell^2/d^{\beta_0}$ goes to 0 too slowly as $d \uparrow \infty$ and leads to candidates that are rejected with increasing frequency as d grows (the acceptance rate thus converges towards 0); see Roberts and Rosenthal (1998). By comparison, RWMH accomplishes its exploration in $\mathcal{O}(d)$ iterations at best, that is whenever it uses a proposal variance satisfying $\sigma_d^2 = \ell^2/d$ ($\beta_0 = 1$). Asymptotically as $d \uparrow \infty$, MALA is thus more efficient than RWMH as it achieves longer steps than its counterpart, leading to a more timely exploration of $\mathcal{S}$.

Under regularity conditions similar to those mentioned in Roberts et al. (1997) and Roberts and Rosenthal (1998) for demonstrating the above results about

RWMH and MALA, we now present theoretical results that lead to a better understanding of the annealed MALA. We underline the fact that the theory expounded in this section is obtained in an asymptotic context, meaning that the conclusions of the theorems introduced are valid for infinite-dimensional target distributions.

Let π be a target density on $\mathbb{R}^d$, with independent and identically distributed (i.i.d.) components as follows

$$\pi(x) = \prod_{i=1}^{d} f(x_i) = \prod_{i=1}^{d} \exp\{l(x_i)\} \ ,$$

where $l(x) = \log\{f(x)\}$. We impose the following regularity conditions on f: all moments of f are bounded; l belongs to C^m (the class of continuous, m-time differentiable functions) for some $m \in \mathbb{N}$; l and its first m derivatives are bounded by a polynomial function, that is $|l(x)|, |l^{(i)}(x)| \le M(x)$ for $i = 1, \dots, m$, where $M(x)$ is a positive polynomial function. We finally assume that $l'(x) = (\log f(x))'$ is a Lipschitz function.

Following the literature on optimal scaling, it is assumed here that the target has a product form, which is quite an important limitation in practice. Indeed, the targets we face usually have at least some level of correlation among their components. In a Bayesian statistical context for instance, the resulting posterior is more than likely to violate the independence assumption, regardless of how 'nice' the initial context might be. Nonetheless, the available asymptotic results usually are relatively robust to the form of the target and, in most cases, are the only available ones on which to rely anyway. Users should however be aware that, generally, the stronger is the correlation among target components, the smaller is the optimal acceptance rate when based on a proposal with independent components (if we can mimic target covariances in the proposal distribution however, then original optimal tuning results typically hold).

Theorem 2 *Consider a target distribution Π whose density satisfies the above conditions with $m = \infty$. Suppose that X_0 is distributed according to the stationary distribution Π. Using a Metropolis-Hastings algorithm with proposal density $q_{\gamma,\sigma}$ as in* (5) *with $\gamma \in (1, 2]$, we find that $\beta_0 = 1$.*

Proof The proof of this result may be found in §B.1 of Appendix B or in §2.4 of Boisvert-Beaudry (2019).

This result says that **for a fixed** $\gamma \in (1, 2]$, the annealed MALA should use a step size parameter of the form $\sigma_d^2 = \ell^2/d$. From Roberts and Rosenthal (1998), this same parameter should be set to $\sigma_d^2 = \ell^2/d^{1/3}$ when $\gamma = 1$. It is thus natural to realize that MALA, which is an approximation to the locally-balanced context, is the best option for infinite-dimensional targets as it explores its space according to $\mathcal{O}(d^{1/3})$. As soon as one departs from the local context ($\gamma = 1$) however, it becomes necessary to generate less variable candidates ($\sigma_d^2 = \ell^2/d$) to avoid facing a null acceptance rate in targets featuring an increasingly large number of dimensions. This of course results in a lengthier exploration of the state space, which is achieved in $\mathcal{O}(d)$ iterations.

Naturally, as $\gamma \downarrow 1$, the behaviour of the annealed MALA sampler should approach that of MALA. So how can the exploration of $\mathcal{S}$ be $\mathcal{O}(d)$ for an arbitrarily small interpolation parameter ($\gamma = 1.001$ say) while it is $\mathcal{O}(d^{1/3})$ for $\gamma = 1$? As it turns out, this discrepancy is corrected through the optimal value for ℓ^2, which becomes infinitely large as $\gamma \downarrow 1$.

Theorem 3 *Consider a target distribution Π whose density satisfies the above conditions with $m = 8$. Suppose that X_0 is distributed according to the stationary distribution Π. Using a Metropolis-Hastings algorithm with proposal density $q_{\gamma,\sigma}$ as in* (5) *with* $\gamma \in (1, 2]$ *and* $\sigma_d^2 = \ell^2/d$ *with* $\ell > 0$*, we find that the asymptotically optimal value of ℓ (as d $\uparrow \infty$) is* $\hat{\ell}_\gamma = 2.38/\{(\gamma - 1)\sqrt{\mathbb{E}[\{l'(X)\}^2]}\}$. *This gives rise to an asymptotically optimal acceptance rate of 0.234.*

Proof The proof of this result is very similar to that of other optimal scaling results in the literature. For the sake of brevity, only the broad lines of this demonstration are presented in §B.2 of Appendix B.

For someone familiar with the optimal scaling literature, the first thing that comes to mind probably is the similarity between the asymptotic behaviours of RWMH and annealed MALA with $\gamma = 2$, as both feature the same $\hat{\ell}$. We however emphasize that a sampler with $\gamma = 2$ is to be used in very low-dimensional settings only (1 or 2 dimensions); as a consequence, these asymptotic results cannot be trusted to accurately tune the sampler. At this point, the finite-dimensional tuning guidelines in Figure 4 of Roberts and Rosenthal (2001), which broadly state to accept 45% of candidates in a one-dimensional RWMH, are likely to be closer to optimality than the asymptotic 23.4%.

For a fixed γ value and a smooth-enough target with sufficiently large d, this theorem may appear as a route towards optimal tuning for the annealed MALA; these are not, however, the key findings from Theorems 2 and 3. What these theorems do tell us is that $\hat{\ell}_\gamma$ becomes arbitrarily large as γ gets closer to 1, which makes up for the fact that σ_d^2 is scaled by a factor of d. Hence, the conclusions of the previous theoretical results should not be understood as negative for the annealed MALA, but rather interpreted as a theoretical validation of our initial intuition from the local and global equilibrium theory of Zanella (2020): *the interpolation parameter γ should decrease in d so as to obtain asymptotic results that are consistent with those of MALA whenever γ approaches 1.* For example, if we let $\gamma_d = 1 + d^{-1/3}$ and then evaluate $\sigma_d^2 = \ell^2/\{(\gamma_d - 1)^2 d\}$, we find that the proposal variance behaves according to $\ell^2/d^{1/3}$, which is in agreement with the theory of Roberts and Rosenthal (1998).

Asymptotics of the annealed MALA Now better equipped to understand the limiting behaviour of the sampler, let $\sigma_d^2 = \ell^2/d^{1/3}$ and $\gamma_d = 1 + d^{-1/3}$. For a target Π whose density satisfies the above conditions with $m = 8$ and a Markov process that starts in stationarity, Theorems 1 and 2 of Roberts and Rosenthal (1998) hold. In particular, the speed at which $\mathcal{S}$ is explored is $\mathcal{O}(d^{1/3})$ and is maximized at the unique value $\hat{\ell}$ for which the annealed MALA accepts 57.4% of candidates. The proofs are practically identical to those in Roberts and Rosenthal (1998) and are omitted.

Although based on asymptotic arguments, Roberts and Rosenthal (1998) observe that their theorems are quite robust to dimensions, especially for symmetrical target densities. The picture is not as clear for asymmetrical ones, where optimal acceptance rates are seen to be closer to 40% in small dimensions. Nonetheless, in their examples, the 57.4% acceptance rate still leads to a relative efficiency in excess of 0.95.

According to our own experiments with the annealed MALA, adjusting the interpolation parameter as a function of the dimension does not appear to negatively impact the robustness of the 57.4% acceptance rate. In the simulation studies of §3.3, the empirical optimal acceptance rate will be seen to gravitate around this value most of the time. With the more complex targets of §3.4, we observe slightly lower optimal acceptance rates (mainly between 40% and 50%). This is to be expected as the assumptions on the form of the target are more strongly violated; the asymptotic results are therefore not as well suited to the context studied, which generally translates into smaller acceptance rates.

Now, if we focus on specific target distributions and study the optimal acceptance rate across γ values (while d is fixed), we observe some cases where the 57.4% rate conveniently holds over the whole range $\gamma \in [1, 2]$. In other cases, the optimal rate starts decreasing when γ is too greatly overestimated with respect to d, eventually reaching an optimal rate of about 40%. Examples of these behaviours will be presented in §4.2. In the real data examples of §4, we shall be tuning our annealed MALA to accept approximately 57% of candidates, so as to be in accord with the theory exposed.

3.2 Tuning the interpolation parameter

We now turn to the tuning of γ_d with respect to dimension. As $d \uparrow \infty$, asymptotic results establish the optimal combination of parameters to be $\beta_0 = 1/3$ and $\gamma = 1$, with $\sigma_d^2 = \ell^2/d^{\beta_0}$ tuned to accept 57.4% of candidates. In practice however, dimensions never are infinitely large and while simulation studies suggest the optimal acceptance rate to be quite robust with respect to d, this does not provide information about the individual tuning of γ_d and σ_d^2 in finite dimensions. Indeed, while the regular MALA only has one parameter to adjust, the annealed MALA has different pairs (γ_d, σ_d^2) leading to the same acceptance rate. In that context, it is thus convenient to propose guidelines for judiciously selecting the interpolation parameter γ_d in terms of d. In fact, it would be interesting to classify dimensions into distinct regimes that we designate as local ($\gamma = 1$), intermediate ($1 < \gamma < 2$), and global ($\gamma = 2$).

We have established, both intuitively and formally, that the interpolation parameter γ_d has to be a decreasing function of the dimension. In particular, we mentioned that a step size $\sigma_d^2 = \ell^2/d^{1/3}$ together with an interpolation $\gamma_d = 1 + d^{-1/3}$ lead to an asymptotically performant sampler, in the sense that exploring $\mathcal{S}$ requires $\mathcal{O}(d^{1/3})$ iterations. Naturally, the speed at which γ_d goes to 1 must be sufficiently rapid for the algorithm to reach the desired asymptotic behaviour. If $\gamma = 2$ for instance, this forces us to use the scaling $\sigma_d^2 \propto d^{-1}$, leading to an exploration in $\mathcal{O}(d)$ steps. Conversely, if the decay is too rapid (the extremal case being $\gamma = 1$ say), then the usual MALA kicks in early on and we do not take advantage of what Zanella (2020)'s theory has to offer (i.e. we find ourselves directly in the local regime even though d is low). We know from the asymptotics of the annealed MALA in §3.1 that $\sigma_d^2 \propto d^{-1/3}$ is the largest proposal variance leading to a positive expected acceptance rate for all $d \geq 1$, when paired to $\gamma_d = 1 + d^{-1/3}$. Now, what about γ_d? Does it decrease from 2 to 1 as slowly as it can, or could we slow its trajectory further down and still converge to a MALA as $d \uparrow \infty$? To answer this question, let

$$\gamma_d = 1 + d^{-\lambda_d} \ , \tag{6}$$

where $\lambda_d : \mathbb{N} \to \mathbb{R}^+$ is a positive function of the dimension. The role of λ_d is similar to that of β_0: it regulates the rate at which γ_d goes to 1 as d increases. Contrarily to the constant β_0 however, λ_d may vary with d; this allows more flexibility, as called upon by the simulation results that will be presented in §3.3.

Theorem 4 *Consider a target distribution Π whose density satisfies the conditions of §3.1 with $m = 8$. Suppose that X_0 is distributed according to the stationary distribution Π. Using a Metropolis-Hastings algorithm with proposal density $q_{\gamma,\sigma}$ as in (5), where γ_d is as in (6) and $\sigma_d^2 = \ell^2/d^{1/3}$ with $\ell > 0$, we find that* $\lim_{d\to\infty} \mathbb{E}[\alpha(X,Y)] > 0$ *if, and only if,* $\lim_{d\to\infty} \lambda_d \geq 1/3$.

Proof The proof of this result may be found in §B.3 of Appendix B or in §3.3 of Boisvert-Beaudry (2019).

This theorem claims that regardless of the rate at which γ_d initially decreases, it must eventually come close to $1 + d^{-1/3}$ in the limit. In setting $\lim_{d\to\infty} \lambda_d = 1/3$, we have access to the largest possible biasing of the proposal mean that still leads to a positive expected acceptance rate. This behaviour shall be observed in the numerical examples of §3.3, where γ_d will be seen to decrease rapidly initially (up to $d \approx 35$) and slowly thereafter (approaching $1 + d^{-1/3}$).

Finding the optimal form for λ_d is arduous as the set of possible functions is vast and the optimal function likely depends on the target under study. Nevertheless, it would be convenient to offer a range of dimensions over which the annealed MALA should be preferred to MALA; in other words, for which d does $\gamma > 1$ perform better than $\gamma = 1$? We propose to numerically study normal targets and recommend a baseline that could eventually be applied, if not exactly, at least approximately to other target distributions. In the next section, we consider two different target distributions and study how the optimal γ_d evolves as a function of d. As will soon be seen, MALA is generally outdone by its newer version with $\gamma > 1$. In §3.4, we then perform a similar study on complex targets to better understand the behaviour of the annealed MALA in presence of strong correlation and/or multimodality.

3.3 Numerical explorations - Simple targets

We consider two target densities that easily scale according to dimension; d-dimensional versions of each target are then studied, where d ranges from 1 to 1,000. Of interest is to gain some insight about the optimal value of the interpolation parameter γ_d as a function of d. To this end, we use the average squared jumping distance (ASJD) of the Markov process as a relative measure of performance for the sampler, $\frac{1}{N}\sum_{t=0}^{N-1} \|X_{t+1} - X_t\|^2$, with N the number of iterations performed.

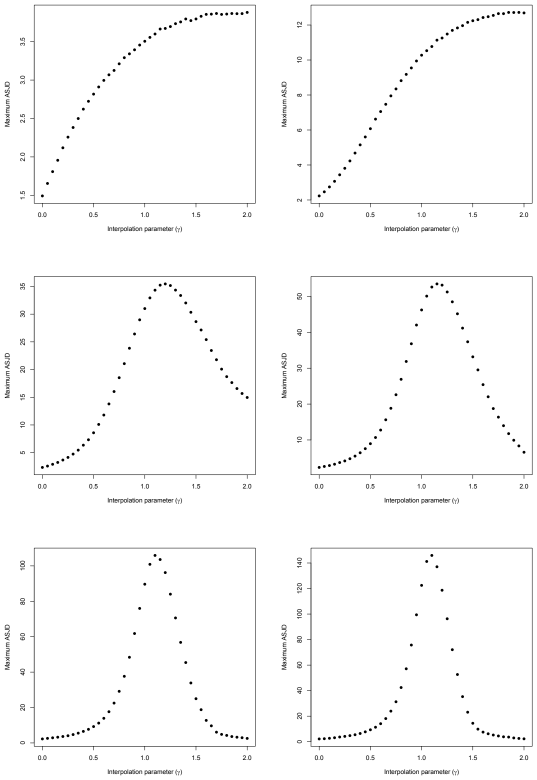

Fig. 1 Bimodal target: max. ASJD against γ for $d = 1, 5, 50, 100, 300, 500$ (left to right)

The first target consists in d independent components, each component being a mixture of two normal distributions such that $X_i \sim 0.6\mathcal{N}(\mu, 1) + 0.4\mathcal{N}(-\mu, 1)$ for $i = 1, \ldots, d$ with $\mu = 1$. The modes of the mixture distribution are not too distant, which is important as MALA-type samplers are not designed to be performant on targets featuring isolated modes.

Using an annealed MALA (aMALA) with proposal as in (5), we obtain samples of size $N = 200,000$ from this target. For a given d, we perform runs using each combination of the parameter values $\gamma \in \{0, 0.05, \ldots, 2\}$ and $\sigma_d^2 = \ell^2/d^{1/3}$ with $\ell^2 \in \{0.1, 0.3, \ldots, 12\}$. For each γ, we then select the proposal variance σ_d^2 that maximises the average squared jumping distance and produce a graph of this maximum distance against γ; see Figure 1. Although aMALA has been defined earlier with $\gamma \in [1, 2]$, we still include $\gamma \in [0, 1)$ in our study; besides gaining a better understanding of the sampler for values of γ outside the usual range, this also allows comparing its performance with that of RWMH ($\gamma = 0$).

The first observation is that below $d = 5$, the global regime is to be favoured. Even if σ_d^2 is not approaching ∞, this parameter is as large as it may be (by comparison to sampling from higher-dimensional versions of the same target). The intermediate regime kicks off after that and at $d = 35$, the optimal γ-value has decreased to 1.25. The drop becomes slower from there, which confirms that λ_d is not constant; for instance, the optimal γ-value is 1.15 at $d = 100$. It is worth noting that we never enter the local regime: the optimal γ-value remains at 1.1 from $d = 300$ to 1,000. The annealed MALA thus seems appropriate even in relatively large dimensions; performance comparisons with the regular MALA are presented below. Finally, although difficult to see from the graphs, we note that the ASJD remains positive for all γ as d grows: when $\gamma = 2$, it is around 2.4 for $d = 300$ and 2.2 for $d = 500$ (with acceptance rates in the range 25-30% in both cases). This illustrates that the asymptotic results of Theorem 3, which state that the annealed MALA explores $\mathcal{S}$ in $\mathcal{O}(d)$ iterations when γ is constant (i.e. not d-ajusted), are attained more rapidly for large γ values. In fact, according to the graphs, the performance of aMALA with $\gamma = 2$ becomes very similar to that of RWMH as d grows.

The second target distribution studied is a $\mathcal{N}(\mathbf{0}, A_d)$, where $\mathbf{0} = (0, \ldots, 0)^\top$ and $A_d = \text{diag}(\tau_1, \ldots, \tau_d)$ with $\tau_i \sim \mathcal{U}\text{nif}(0.5, 2)$, $i = 1, \ldots, d$. The d target components are thus independent, with uniformly distributed individual variances. Figure 2 presents the results obtained from the same experiment as before. This time, the global regime is valid for a few dimensions only and at $d = 5$, we comfortably find ourselves in the intermediate regime ($\gamma = 1.85$). By the time we reach $d = 35$, the optimal γ is 1.35 (by opposition to 1.25 for the previous target). Once again, the optimal value for the interpolation parameter decreases much more rapidly in small dimensions than it does in higher ones. The optimal value is 1.2 at 300, then drops to 1.15 at $d = 500$ and remains there until at least $d = 1,000$.

The optimal γ values are not exactly the same in both examples, but their decline rates appear to follow a similar pattern. We also note that the local regime has yet to be attained in either case. In the bimodal context, a quick calculation using γ_d shows that λ_d exceeds 1/3 from $d = 25$, and then slowly decreases towards 1/3 as $d \to \infty$. The normal target studied in the second example does not meet the requirements imposed by Theorem 4, as its components are not i.i.d; in that example, λ_d approaches 1/3 from below as d grows. In both cases, λ_d is not constant; it initially increases rapidly (corresponding to a rapid decay of γ_d up to $d \approx 35$), then slowly approaches 1/3 (either from above or below). Interestingly, a rule of thumb for λ_d seems to emerge from our numerical explorations; these investigations, which included asymmetrical targets (such as the above-mentioned mixture of normals) and weakly correlated ones, revealed that the empirical λ_d are fairly well approximated by $\lambda_d^* = \min(\sqrt{d/100}, 1/3)$. Figure 3 illustrates the relationship between d and γ_d for the examples discussed above and presents the approximation $\gamma_d^* = 1 + d^{-\min(\sqrt{d}/10, 1/3)}$ as a function of d.

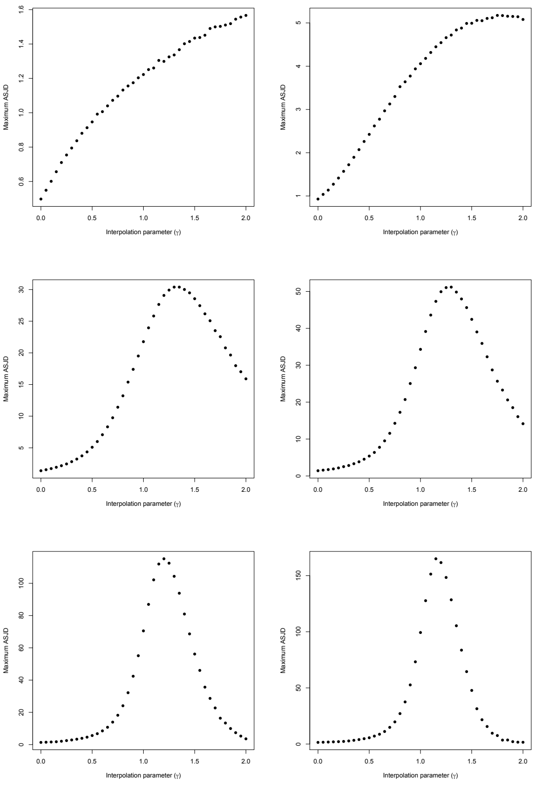

Fig. 2 Normal target: max. ASJD against γ for $d = 1, 5, 50, 100, 300, 500$ (left to right)

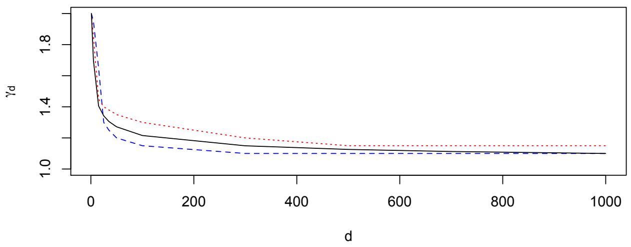

Fig. 3 Interpolation parameter as a function of target dimension: bimodal (dashed blue), normal (dotted red), approximation λ_d^* (solid black)

Knowing that the annealed MALA leads to an improved exploration of $\mathcal{S}$ by comparison to the usual one, it is now worth wondering if the promised benefits are worth tuning an extra parameter. Table 1 presents the efficiency gains that are available from introducing an interpolation parameter γ, for both targets previously studied. The improvement with respect to MALA, measured in percentage, is defined as

$$\% \text{ improvement} = \frac{\text{ASJD}_{\gamma_{opt}} - \text{ASJD}_{\gamma=1}}{\text{ASJD}_{\gamma=1}} \times 100\% \ .$$

The table also records, for each d, the empirical acceptance rates that correspond to the maximum ASJD obtained.

The main feature emerging from Table 1 is that the benefit from implementing an annealed MALA does not necessarily vanish as γ gets closer to 1; on the contrary, it seems to become more significant as d increases. For the normal target, this improvement is over 50% from $d = 100$. For the mixture target, it increases from 15% to 25% as the dimensions go from 100 to 1,000. These efficiency gains, which are available at no additional computational cost, are thus present across all d but are not necessarily proportional to γ_d. This is not overly surprising as in the limit, the sampler behaves according to a diffusion process with speed measure $\upsilon(\ell, \gamma = 1)$; the optimal tuning $\hat{\ell}$ is therefore the value that maximizes the speed $\upsilon(\ell)$. If we make the dependence on $\gamma = 1$ explicit, the speed measure satisfies $\upsilon(\ell) = 2\ell^2 \Phi(-\ell^3 K/2)$, with $K = (3\gamma/2 - 1)\{\mathbb{E}[3l'''(X)^2 - 3\gamma^2 l''(X)^3]/12\}^{1/2}$. Now imagining that γ could vary in that equation, and that ℓ would then be adjusted so as to optimize the speed measure, it is easy to see that the impact on $\upsilon(\ell)$ would not be linear; similar outcomes are then expected in finite dimensions. As prescribed by the theory, these gains will eventually decrease towards 0 as d continues to grow but in practice, the point where MALA becomes optimal has not been reached.

Another particularity of Table 1 is that the empirical acceptance rates recorded are over 80% for the pair $(d = 1, \gamma = 2)$. Since a high proportion of candidates are being accepted, this indicates that the large γ has to be balanced by a relatively small step size σ. As σ simultaneously affects the proposal variance and bias, this then leads to a moderate benefit in terms of ASJD (compared to larger d). As d grows, acceptance rates are seen to decrease slightly and then rise again to stabilize around 57%. For the normal target, rates vary between 52% and 67% for d ranging from 5 to 1,000. For the bimodal target there is more instability for $5 \leq d \leq 35$, but after that rates fluctuate between 50% and 56%. Overall, a rate of 57% appears as the safest tuning choice for $d \geq 5$, while one could afford to be more aggressive with very low-dimensional targets ($d < 5$).

3.4 Numerical explorations - Complex targets

The standard MALA has some documented issues when it comes to exploring strongly anisotropic target distributions. Naturally, we do not expect the annealed MALA to offer a miraculous solution where MALA struggles or fails. Nevertheless, it is of interest to study the behaviour of our sampler in various contexts, and to understand which γ to favour with such targets.

In cases where π has strongly correlated components, which is one way anisotropy may arise, MALA is known to require a very small tuning parameter σ^2. Indeed, using the gradient of the log-target to bias our trajectory independently in each direction might be misleading; a small σ^2 thus compensates for the fact that we overlook correlation among target components.

In a similar fashion, we do not expect the annealed MALA, in its actual expression at least, to do much better than MALA. The annealed MALA proposes candidates that feature a stronger bias than its version with $\gamma = 1$, hence we expect even smaller values of σ^2 to be required. In correlated contexts, the annealed MALA with $\gamma \in (0, 1)$ might even perform better than $\gamma \in [1, 2]$; indeed, small γ still guide the direction of the next candidate, which is better than going blindly, but without being too aggressive in a given direction. While this is, in part, what we will observe in the first example below, we will also realize that performances for all $\gamma \in [0, 2]$ are rather weak.

It turns out that the first-order approximation of the biasing function g in §2.4 is not precise enough to capture the characteristics of complex targets. To overcome this problem, we simply approximate g using a second-order Taylor expansion; this of course implies compromising on the simplicity of the proposal distribution, but appears necessary for an efficient exploration of the state space. Developing $g(x, y) \propto \{\pi(y)\}^{\gamma/2}$ with respect to y and around x leads to

$$\{\pi(y)\}^{\gamma/2} \approx \exp\left\{\frac{\gamma}{2}\log\pi(x) + \frac{\gamma}{2}\nabla\log\pi(x)(y - x) + \frac{\gamma}{4}\nabla^2\log\pi(x)(y - x)^2\right\} \ .$$

The approximated biased proposal density then becomes

$$\begin{aligned} &q_{g,\sigma}(x, y)\mathrm{d}y \\ &\propto \exp\left\{\frac{\gamma}{2}\nabla\log\pi(x)(y - x) + \frac{\gamma}{4}\nabla^2\log\pi(x)(y - x)^2\right\} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma^2}(y - x)^\top(y - x)\right\}\mathrm{d}y \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\left(y - x - \frac{\gamma\sigma^2}{2}A(x)\nabla\log\{\pi(x)\}\right)^\top \right. \\ &\qquad \left. A^{-1}(x)\left(y - x - \frac{\gamma\sigma^2}{2}A(x)\nabla\log\{\pi(x)\}\right)\right\}\mathrm{d}y \ , \end{aligned}$$

with $A(x) = \{I_d - \gamma\sigma^2\nabla^2\log\pi(x)/2\}^{-1}$. The resulting proposal distribution is thus a position-dependent

	Bimodal target			Normal target		
d	γ_{opt}	% improv.	Acc. rate (%)	γ_{opt}	% improv.	Acc. rate (%)
1	2	11.39	80.96	2	28.66	86.80
5	1.95	25.09	54.28	1.85	27.10	66.90
15	1.65	20.18	34.66	1.45	26.33	60.99
25	1.3	13.07	41.29	1.4	34.20	52.17
35	1.25	12.66	47.78	1.35	33.63	59.96
50	1.2	14.08	50.48	1.35	40.13	58.77
100	1.15	15.29	55.76	1.3	49.71	61.83
300	1.1	18.06	55.34	1.2	63.88	58.75
500	1.1	18.95	54.81	1.15	67.06	61.64
700	1.1	22.52	54.50	1.15	72.88	60.35
1000	1.1	25.04	53.45	1.15	72.39	54.47

Table 1 Simple targets: % improvement in terms of ASJD, annealed MALA vs usual MALA

conditioned annealed MALA

$$Y_{t+1} \sim \mathcal{N}\left(X_t + \frac{\gamma\sigma^2}{2}A(x)\nabla\log\{\pi(X_t)\}, \sigma^2 A(x)\right) ,$$

where $A(x)$ is a conditioning matrix that depends on the current state of the process. When $\gamma = 1$, this sampler is a special case of the manifold MALA of Girolami and Calderhead (2011).

The role of the conditioning matrix $A(x)$ consists in capturing the target correlation structure. Upon examination, $A^{-1}(x)$ involves a multiple of $-\nabla^2 \log \pi(x)$, to which we add the identity matrix I_d. In this version of the sampler, the term $\gamma\sigma^2/2$ is then responsible for calibrating the weight of the negative Hessian matrix (of the log target density) so as to obtain a positive definite matrix $A(x)$. With log-concave target densities, this characteristic is automatically satisfied and it is therefore preferable to simply let $A^{-1}(x) = -\nabla^2 \log \pi(x)$. In other cases however, the identity matrix is not only required, but might even be largely predominant to ensure that the conditioning matrix is positive definite for every $x \in \mathcal{S}$. When this happens, the resulting sampler compares to the aMALA with independent components, obtained using a first-order Taylor expansion of $g(x,y) \propto \{\pi(y)\}^{\gamma/2}$. In absence of log-concavity, it is thus more appropriate to use a positive-definite approximation of $-\nabla^2 \log \pi(x)$ for $A^{-1}(x)$.

Given that $A(x)$ succeeds in capturing the target correlation structure reasonably well, we expect this conditioned annealed MALA to behave as before, that is to favour large γ in small dimensions and values closer to 1 in large dimensions. To illustrate this, consider a d-dimensional hierarchical target with $X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \Gamma(3,1)$, and $X_i|X_{1:2} \sim t_7(X_1, 1/\sqrt{X_2})$ for $i = 3,\dots,d$; the density of a generalized Student-$t_\nu(\mu,\eta)$ is proportional to $[1+\{(x-\mu)/\eta\}^2/\nu]^{-(\nu+1)/2}$. The components X_1, X_2 respectively act through the location and scale of the variables X_i while the Student distribution destroys conjugacy and, along with it, any niceness that could result from working exclusively with normal distributions.

We perform a simulation study similar to that of §3.3. We run $N = 200,000$ iterations of an annealed MALA with proposal as in (4) for various combinations of d, σ_d^2, and $\gamma \in [0,2]$. For each d, we record the values of γ and σ_d^2 that maximize the standardized ASJD. The latter is similar to the ASJD, but is more appropriate for anisotropic target distributions and is defined as

$$\text{stand. ASJD} = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{d}\frac{1}{\omega_i^2}\left(X_{i,t+1} - X_{i,t}\right)^2 ;$$

ω_i^2 is the marginal variance of the ith target component.

Due to the strong correlation between target components, isotropic versions of the annealed MALA (including RWMH and MALA) do not succeed in efficiently exploring the space. Table 2 reports the maximum standardized ASJD and corresponding acceptance rate obtained for a selection of dimensions. Because the sampler ignores correlation, it has to be much more conservative in its biasing of the proposal mean when $d < 20$; consequently, $\gamma \in (0,1)$ offer better performances than $\gamma \in [1,2]$ in small dimensions. As d grows however, we find ourselves forced to propose increasingly conservative moves to keep the acceptance rate from going to 0 (σ_d^2 is decreasing in d); candidates then become conservative to a point where the correlation structure does not matter anymore and it becomes preferable to set γ closer to 2.

Admittedly however, none of the above isotropic samplers offers a convincing performance and it appears necessary to implement a conditioned version of the annealed MALA. Since the t distribution is not log-concave, $-\nabla^2 \log \pi(x)$ is not positive definite everywhere; we instead use the conditioning matrix A with $A^{-1} = -\mathbb{E}[\nabla^2 \log \pi(X)]$, where the expectation is with respect to π. We could alternatively use a position-dependent conditioning matrix $A(x_{1:2})$ with $A^{-1}(x_{1:2}) =$

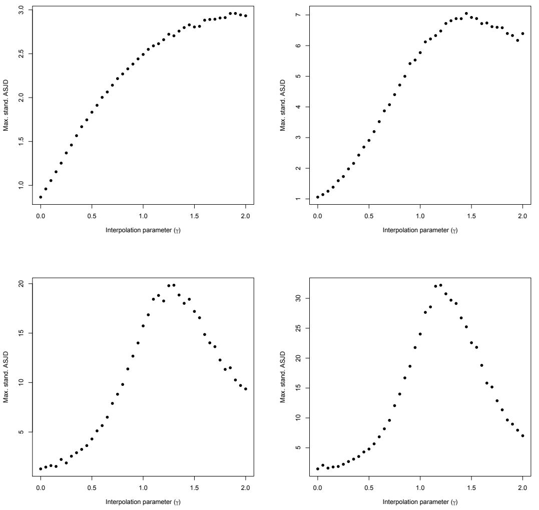

Fig. 4 Hierarchical target: max. standardized ASJD against γ for $d = 3, 12, 52, 102$ (left to right)

$-\mathbb{E}[\nabla \log \pi(x_{1:2}, X_{3:d})]$, where the expectation is with respect to the conditional distribution of $X_{3:d}|x_{1:2}$, but we opt for the former. With the matrix A, the results reported in Table 2 and the graphs in Figure 4 are very similar to those obtained in §3.3; in particular, relying on γ_d^* (2nd column of Table 2) would yield results that are extremely close to optimal.

The case of multimodal target distributions has its own particularities. In order to isolate these special features from the effects of correlation, we first consider a bimodal target with independent components, $X \sim pt_\nu(\mu, I_d) + (1-p)t_\nu(-\mu, I_d)$, with $\mu = (\mu_1, 0, \ldots, 0)$ and $\mu_1 > 0$. The gradient of the target log-density,

$$\nabla \log \pi(x_t) \\ = p^*(x_t)\nabla \log \pi_1(x_t) + \{1 - p^*(x_t)\}\nabla \log \pi_2(x_t) \ ,$$

consists in a weighted average of the distinct gradients with weight function $p^*(x_t) = p\pi_1(x_t)/\{p\pi_1(x_t) + (1-p)\pi_2(x_t)\}$; here, π_1 and π_2 implicitly represent the Student densities involved in the mixture.

Suppose temporarily that $d = 1$, $\nu = 7$, $p = 0.5$, and $\mu_1 = 2$; the resulting one-dimensional target is symmetrical about 0, with distinct modes at 2 and -2, respectively. When the process is in the target's left tail, $\nabla \log \pi \approx \nabla \log \pi_2$ pulls the process up the left mode, towards the center of the distribution (and similarly for the right tail). When the process is anywhere between the modes however, $\nabla \log \pi_1$ and $\nabla \log \pi_2$ pull in different directions, resulting in a weaker bias than with unimodal distributions. Therefore, while a large value of γ definitely is desirable in the tails, the approach to favour is not as clear when the process is located between modes; in particular, should we opt for a strong biasing of the proposal mean (large γ) or simply trust a comparatively larger proposal variance (small γ) to readily explore the state space?

It turns out that both approaches have their own merits. If we fix $p = 0.5$, $d = 1$, and let μ_1 increase from 0 to 15, we find that $\gamma = 2$ leads to an optimal version of the annealed MALA in terms of ASJD, as expected. As the modes get farther away however, the efficiency curve (ASJD vs. γ, not included) becomes flat at the top, depicting a plateau that widens with the distance between modes, and indicating that several γ yield similar performances. When the modes become too distant, the plateau eventually narrows again and it becomes preferable to rely on $\gamma = 2$ only. In that case, relying on a large bias directed towards the alternate (distant) mode is a safer bet than trusting the variance to achieve a mode change. We however note that the associated acceptance rate becomes very small, which indicates that a specialized, regional sampler might be more appropriate with distant modes.

Now, as the dimension d increases, we have to make sure that both modes are visited, regardless of the distance between them. For an appropriate exploration of the state space, we precondition according to the marginal variances $(\frac{\nu}{\nu-2}+\mu_1^2, \frac{\nu}{\nu-2}, \ldots, \frac{\nu}{\nu-2})$. As d grows, the plateau previously observed gradually vanishes since the bimodal component then represents only one of several other components in the standardized ASJD; this also holds for various weights p.

Now, let us rotate our initial bimodal distribution: we use $\mu = (\mu_1, \ldots, \mu_1)/\sqrt{d}$, leading to a constant distance between modes for all d. This is the same target as before, but positioned differently on the sample space. One notable implication of this transformation is that it yields a target with correlated components; the covariance matrix consists in $\frac{\nu}{\nu-2} + \frac{\mu_1^2}{d}$ on the diagonal and $\frac{\mu_1^2}{d}$ off the diagonal. Using an approximation A to this covariance matrix in the preconditioned annealed MALA obviously leads to results that are consistent with those just described in the independent case.

We pursue our study using a more general target,

$$X \sim pt_\nu(\mu, \Sigma_1) + (1-p)t_\nu(-\mu, \Sigma_2) \ ,$$

where $\mu = (\mu_1, 0, \ldots, 0)$ and $\mu_1 > 0$. We fix $p = 0.5$, $\nu = 7$, $\mu_1 = 1.5$, $\Sigma_1 = diag(1, 2, 1, \ldots, 2)$, and similarly $\Sigma_2 = diag(2, 1, 2, \ldots, 1)$. In two dimensions for instance, the Student distributions are spread out in directions that are orthogonal to each other (T shape). To obtain a decent performance on such a complex target, we use a preconditioning matrix A; since the negative Hessian of the log target is not positive definite everywhere, we simply use A with $A^{-1} = p^*(x_t)\Sigma_1^{-1} +$

		Isotropic aMALA				Preconditioned aMALA			
d	γ_d^*	γ_{opt}	stand. ASJD	% improv.	Acc. rate (%)	γ_{opt}	stand. ASJD	% improv.	Acc. rate (%)
3	1.83	0.70	1.13	5.39	44.91	1.90	2.96	18.79	58.33
12	1.44	0.25	0.58	46.08	14.87	1.45	7.05	22.18	46.92
27	1.33	1.60	0.98	61.96	26.76	1.30	12.34	23.65	43.75
52	1.28	2.00	1.27	94.93	59.69	1.30	19.85	26.19	39.71
102	1.21	1.70	1.17	82.55	45.44	1.20	32.20	34.00	44.46
502	1.13	1.90	0.68	54.69	66.57	1.10	109.93	38.10	40.05

Table 2 Hierarchical target: optimal performances of isotropic and preconditioned annealed MALA, per dimension d, when $\gamma \in [0, 2]$. We record the improvement, in terms of standardized ASJD, of optimal annealed MALA vs usual MALA ($\gamma = 1$)

	Preconditioned aMALA with γ_d^*				Preconditioned aMALA with γ_{opt}			
d	γ_d^*	stand. ASJD	% improv.	Acc. rate (%)	γ_{opt}	stand. ASJD	% improv.	Acc. rate (%)
2	1.91	2.23	4.64	45.87	2.00	2.26	5.72	48.61
10	1.48	7.80	25.69	45.34	1.70	7.88	26.98	44.00
25	1.34	13.53	35.57	56.87	1.50	14.18	42.13	49.99
50	1.27	10.69	19.95	73.04	1.20	11.09	24.41	76.94
100	1.22	6.31	24.54	74.74	1.20	6.31	24.54	74.74
200	1.17	3.67	21.67	85.37	1.15	3.67	21.67	85.37

Table 3 Bimodal target: performances of preconditioned annealed MALA, per dimension d, when using γ_d^* and γ_{opt}, respectively. We record the improvement, in terms of standardized ASJD, of the annealed MALA vs usual MALA ($\gamma = 1$)

$\{1 - p^*(x_t)\}\Sigma_2^{-1}$. Table 3 presents efficiency gains, for various dimensions d, obtained using the approximation γ_d^* and the optimal γ_{opt}. In both cases, we record the standardized ASJD, the % improvement over MALA ($\gamma = 1$), and the corresponding % of accepted candidates (note that results corresponding to a rounded version of γ_d^* – to the nearest 0.05 – are reported). These numbers were obtained by running $N = 200,000$ iterations of the annealed MALA on a range of values for σ_d^2; γ_{opt} is the value that corresponds to the largest standardized ASJD in a given dimension.

In two dimensions, we observe the same phenomenon as for the simpler bimodal target: since the distance between modes is considerable, then $\gamma \in (1.3, 2)$ lead to similar performances (within 2% of optimal performance) and the improvement over $\gamma = 1$ is not as sizeable as with unimodal targets. From Table 3, we also see that the progression of γ_{opt} is not as close to γ_d^* as before. Nonetheless, since several γ offer similar performances, the rule γ_d^* still yields almost optimal results. As d grows, these particularities gradually vanish and the usual behaviour prevails.

Because of the high complexity of this specific target, the performance does not hold very well in larger dimensions (the standardized ASJD suddenly starts decreasing). In spite of this, the rule γ_d^* is still worth implementing as it offers improvements in the 5%-40% range. When d grows, the opposing target variances in each mode make it laborious for the MALA and aMALA to explore the space. In fact, for the process to move around in the multidimensional space, it becomes necessary to propose steps that are small to the point where they are almost all accepted; this yields unusually large acceptance rates, which indicate that a different, more specialized sampler might be more appropriate for high-dimensional versions of this target.

By repeating the same exercice as before and rotating this distribution ($\mu = (\mu_1, \ldots, \mu_1)/\sqrt{d}$), we obtain a target with opposing correlation structures in each mode. Using a preconditioning matrix A with $A^{-1} = p^*(x_t)\hat{\Sigma}_1^{-1} + \{1 - p^*(x_t)\}\hat{\Sigma}_2^{-1}$, where $\hat{\Sigma}_1^{-1}$ and $\hat{\Sigma}_2^{-1}$ are estimates of Σ_1^{-1} and Σ_2^{-1}, obviously leads to results that are consistent with those just discussed.

4 Bayesian logistic regression

We now explore the performance of the annealed MALA using different efficiency criteria and real datasets in the context of Bayesian logistic regression. We consider an $n \times d$ design matrix X that contains, for each of the n subjects, information about d explanatory variables (including an intercept for the model). Associated to these variables are d regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{d-1})$ and a response vector $\mathbf{t} \in \{0,1\}^n$ containing n binary variables that indicate whether an event of interest has happened or not. Of interest is to determine

$$p_i = \mathbb{P}(t_i = 1|\boldsymbol{\beta}) \;=\; \frac{1}{1 + \exp\{-\mathbf{X}_i\boldsymbol{\beta}\}}\ ,$$

Name	n	d	Response variable
Pima Indian	532	8	Diabetes (yes/no)
German Credit	1000	25	Credit (good/bad)
Australian Credit	690	15	Credit (approved/refused)
Heart	270	14	Heart disease (yes/no)

Table 4 Details about datasets: Bayesian logistic regression

the probability that the event happens given $\mathbf{X}_i$, the explanatory variables observed for the i-th subject. To estimate the regression parameter $\boldsymbol{\beta}$, we use a Bayesian approach; the likelihood function satisfies $f(\mathbf{t}|\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{t_i}(1-p_i)^{1-t_i}$ and the prior distribution is chosen to be vague, $\beta_j \sim \mathcal{N}(0, 100)$, $j = 0, \ldots, d-1$. We sample from the posterior distribution of $\boldsymbol{\beta}$ using MCMC and then estimate the regression coefficients using the output.

We analyze the four datasets presented in Table 4 and studied in Girolami and Calderhead (2011). These datasets offer various challenges: the dimension of $\boldsymbol{\beta}$ varies between 8 and 25, while n goes from 270 to 1,000. Since d is relatively small in all four cases, an annealed MALA with $\gamma > 1$ seems particularly appropriate. To avoid numerical issues, all variables have been standardized. We refer the interested reader to Michie et al. (1994) and Ripley and Hjort (1996) for more information about these datasets and the context in which the observations have been collected.

4.1 Simulation results

For each dataset, samples from the posterior distribution $\pi(\boldsymbol{\beta}|\mathbf{t})$ are obtained using RWMH ($\gamma = 0$), MALA ($\gamma = 1$), and annealed MALA with $\gamma = 0.3, 0.6, 1.2, 1.4, 1.6, 1.8, 2$; although γ should be in the range $[1, 2]$, we include some γ values in $(0, 1)$ in order to validate our theoretical results. We also provide results based on the approximation $\gamma_d^* = 1 + d^{-\min(\sqrt{d}/10, 1/3)}$. In all cases, the proposal variance σ^2 is tuned so as to approach an average acceptance rate of 23% for the RWMH and 57% for the MALA and annealed MALA. These asymptotically optimal rates are our best option given their robustness to dimension and the lack of tuning guidelines for low-dimensional target distributions.

As was done in Girolami and Calderhead (2011), we perform 10,000 iterations and discard the first 5,000 as burn-in. Samplers' performance is measured using the ASJD and the effective sample size (ESS). For each target component X_i, the latter is computed as $\mathrm{ESS}_i = N/\{1 + 2\sum_k \rho_k\}$, where N is the number of iterations and $\sum_k \rho_k$ is the sum of the K monotone sample autocorrelations as estimated by the initial monotone sequence estimator of Geyer (1992); we then report the median of ESS_i. The ESS may be interpreted as the size of an i.i.d. sample that contains as much information as the current correlated sample; we thus wish to maximize this value. We also record running times in seconds, and report Time/ASJD.

Tables 5 to 8 compile the performance results of the various samplers implemented, for each of the four datasets. Unsurprisingly, RWMH has the lowest per-iteration cost; MALA and its annealed version require up to twice as long to complete the same number of iterations. In all cases, the RWMH performance is however rather weak, whether it be according to ASJD or ESS, and so MALA ends up being more efficient when accounting for computational effort. In particular, the optimal γ values are significantly greater than 1 for all datasets and the benefits that are available from implementing the annealed MALA over the usual one range from 13% (Pima Indian, $d = 8$) to 30% (Australian Credit, $d = 15$). As it turns out, using γ_d^* yields the highest ASJD in 2 cases out of 4 and the highest ESS in 3 cases out of 4. Even when γ_d^* is not optimal among the values implemented, the ASJD it produces is always **at least** 98.5% of the highest ASJD reported.

In all four cases, selecting $\gamma < 1$ leads to significant losses in terms of ASJD and ESS compared to using a regular MALA, which is in agreement with the theory previously developed. While overestimating γ might not be as dramatic and might still produce efficiency gains, overdoing it could also lead to an underperforming sampler compared to MALA. In the examples analyzed, setting $\gamma = 2$ produces a small loss in two cases (Pima Indian and Heart), while setting $\gamma = 1.8$ or 2 leads to more substantial losses in the highest dimensional case (German Credit, $d = 25$). Similar conclusions hold for the ESS and we note that the latter decreases more rapidly than the ASJD.

In general, when γ is too large with respect to d, the biasing term $\gamma\sigma^2\nabla\log\{\pi(x)\}/2$ of the proposal distribution is too aggressive and candidates find themselves in regions of low target density. Since these candidates are likely to be rejected, the proposal variance σ^2 needs to be reduced so as to lower the weight in front of the gradient and maintain an acceptance rate of 57%. This naturally results in small, highly-correlated steps, and therefore small ASJD/ESS. In such cases, if we settle on an acceptance rate smaller than the usual 57% (so a larger σ), then conclusions are different. In the German Credit example, setting $\gamma = 1.8$ and tuning σ^2 so as to obtain an acceptance rate of 42% leads to an ASJD greater than MALA (0.08293 against 0.08049). Therefore, where there is a risk of overestimating γ too

Algorithm	ASJD	Time (sec)	Time/ASJD	ESS	% improvement
RWM	0.01762	2.29	130.10	134.65	-78.95
($\gamma = 0.3$)	0.02125	4.05	190.58	156.17	-74.96
($\gamma = 0.6$)	0.04593	4.06	88.37	338.77	-45.87
MALA	0.08373	4.05	48.39	619.32	0
($\gamma = 1.2$)	0.09214	4.04	43.95	653.42	10.04
($\gamma = 1.4$)	**0.09492**	4.03	42.47	**658.30**	**13.35**
($\gamma = 1.56$)	0.09356	4.07	43.50	631.76	11.74
($\gamma = 1.6$)	0.09239	4.05	43.84	625.81	10.34
($\gamma = 1.8$)	0.08793	4.05	46.10	554.44	5.01
($\gamma = 2$)	0.07902	4.05	51.26	422.82	-5.62

Table 5 Results for the Pima Indian dataset ($n = 532, d = 8$)

Algorithm	ASJD	Time (sec)	Time/ASJD	ESS	% improvement
RWM	0.00920	6.53	709.51	44.61	-88.56
($\gamma = 0.3$)	0.00979	12.26	1252.77	52.66	-87.83
($\gamma = 0.6$)	0.02837	12.24	431.62	139.97	-64.74
MALA	0.08049	11.20	148.20	365.76	0
($\gamma = 1.2$)	0.09662	12.18	126.09	415.36	19.89
($\gamma = 1.34$)	**0.10015**	12.20	121.82	**425.57**	**24.42**
($\gamma = 1.4$)	0.09932	12.13	122.22	424.15	23.24
($\gamma = 1.6$)	0.09112	12.69	139.35	373.59	13.21
($\gamma = 1.8$)	0.01045	12.22	1169.68	53.91	-87.02
($\gamma = 2$)	0.00621	12.09	1946.92	35.44	-92.28

Table 6 Results for the German Credit dataset ($n = 1000, d = 25$)

Algorithm	ASJD	Time (sec)	Time/ASJD	ESS	% improvement
RWM	0.02470	3.94	159.38	98.80	-85.71
($\gamma = 0.3$)	0.02735	6.67	243.88	103.30	-84.36
($\gamma = 0.6$)	0.06958	6.58	94.69	247.27	-60.22
MALA	0.17288	6.44	37.27	621.18	0
($\gamma = 1.2$)	0.20736	6.55	31.59	706.92	19.94
($\gamma = 1.4$)	0.22228	7.17	32.26	761.92	28.57
($\gamma = 1.42$)	0.22269	6.83	29.22	**765.05**	28.81
($\gamma = 1.6$)	**0.22459**	6.53	29.09	724.38	**29.91**
($\gamma = 1.8$)	0.21364	6.68	31.28	659.24	23.57
($\gamma = 2$)	0.18624	6.59	35.43	481.16	7.72

Table 7 Results for the Australian Credit dataset ($n = 690, d = 15$)

greatly, lower acceptance rates (around 40%) seem more appropriate.

4.2 The impact of σ^2

In light of the previous examples, it seems pertinent to examine in more depth the impact of σ^2 on the performance of the sampler. We saw that greatly overestimating γ sometimes leads to an algorithm that does not perform as well as MALA. Could this be explained by the fact that σ^2 was tuned so as to favour the latter? We remind readers that the 57% acceptance rate targeted arises from asymptotic results and is thus more likely to be appropriate when γ is small.

We focus on the Pima Indian and German Credit datasets, as those cases suffered efficiency losses under large γ. For each of these examples, we fix γ and run the annealed MALA with several σ^2 so as to cover a range of acceptance rates going from 0 to 1. For each γ, we present a graph of ASJD against acceptance rate; we note that small σ^2 correspond to high acceptance rates and vice versa. Figure 5 reports graphs related to Pima Indian, while Figure 6 presents those of German Credit.

In the case of Pima Indian ($d = 8$), the optimal acceptance rate is approximately 57% for $\gamma = 1$, which supports the previous claim about robustness of theoretical results for finite-dimensional targets. As γ increases, the optimal acceptance rate does as well, finding itself around 62% for $\gamma = 1.2$ and 70% for $\gamma = 1.4$. The optimal acceptance rate then starts decreasing; it is close to 57% for $\gamma = 1.6, 1.8$, and then lowers to about

Algorithm	ASJD	Time (sec)	Time/ASJD	ESS	% improvement
RWM	0.05264	1.71	32.52	93.06	-85.94
$(\gamma = 0.3)$	0.06272	3.04	48.50	101.88	-83.16
$(\gamma = 0.6)$	0.15287	3.07	20.10	231.57	-58.95
MALA	0.37430	3.01	8.04	577.14	0
$(\gamma = 1.2)$	0.43199	2.99	6.92	626.22	14.97
$(\gamma = 1.4)$	0.45584	2.98	6.54	646.74	21.79
$(\gamma = 1.41)$	**0.45613**	3.05	6.68	**677.27**	**21.86**
$(\gamma = 1.6)$	0.44876	3.02	6.73	586.91	19.89
$(\gamma = 1.8)$	0.41554	3.10	7.45	481.08	11.02
$(\gamma = 2)$	0.35810	3.05	8.53	395.54	-4.33

Table 8 Results for the Heart dataset $(n = 270, d = 14)$

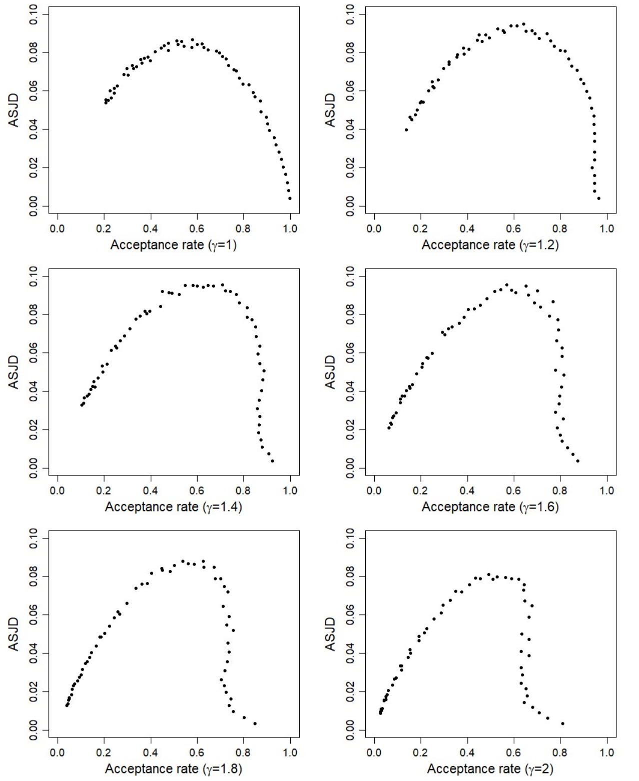

Fig. 5 ASJD vs acc. rate for $\gamma = 1, 1.2, 1.4, 1.6, 1.8, 2$ (left to right), Pima Indian

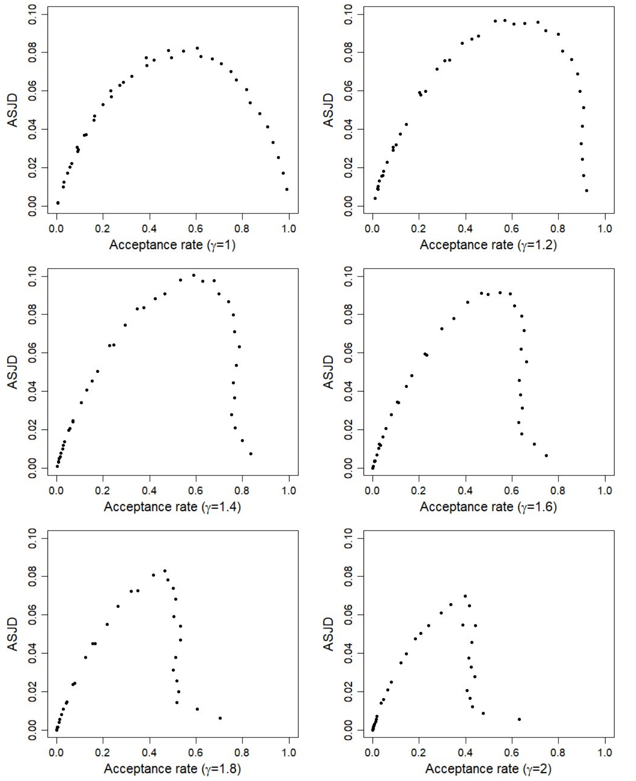

Fig. 6 ASJD vs acc. rate for $\gamma = 1, 1.2, 1.4, 1.6, 1.8, 2$ (left to right), German credit

50% when $\gamma = 2$. The graphs confirm that the optimal γ lies around 1.4, since the maximum ASDJ when $\gamma = 1.4$ is the highest among all curves. Optimally tuned versions of the annealed MALA with $\gamma = 1$ or 2 offer similar performances, with a slight edge for the traditional MALA. The shape of the curves is also evolving: they become steeper as γ increases, which indicates that simultaneously overestimating the interpolation parameter and the acceptance rate is risky in terms of efficiency. Nonetheless, efficiency curves remain fairly flat at the maximum in this 8-dimensional example, so tuning the sampler to accept 57% of candidates leads to an almost optimal sampler, independently of γ.

In the German credit example, the optimal acceptance rate is close to 57% from $\gamma = 1$ to 1.6, falls slightly below 50% for $\gamma = 1.8$, and then further down to 40% for $\gamma = 2$. In Figure 6, we also observe that the annealed MALA with $\gamma = 1.4$ offers an efficiency gain of about 25% compared to MALA. This benefit however gets smaller as γ increases, almost vanishing around $\gamma = 1.8$ and then becoming a 10% loss when $\gamma = 2$. The sampler also becomes more sensitive to the choice of σ^2 as γ grows; in particular, several σ^2 lead to a common acceptance rate but widely different ASJD. As σ^2 decreases, the variability of candidates decreases and so does the drift term $\gamma\sigma^2\nabla\log\{\pi(x)\}/2$. For several σ^2, the effect of a smaller drift seems to be almost perfectly counteracted by a smaller variance, leading to candidates that are neither accepted more nor less often. The decreasing variability of candidates however produces smaller ASJD and this phenomenon amplifies as γ approches 2. In such cases, overestimating the acceptance rate may have a dramatic effect, which is what happened in Table 6 for $\gamma = 1.8$ and 2.

5 Discussion

The usual MALA originates from the discretization of a Langevin diffusion process. Although researchers and users assumed this sampler to enjoy a certain form of optimality due to its ties to the Langevin process, our empirical results suggest that this is only true asymptotically, for *infinite-dimensional* distributions of interest. In fact, the algorithm implemented over the last decades can typically be improved by adding a single tuning parameter, for which simple tuning heuristics exist. This appears as an efficient way to preserve MALA's efficiency through dimensions.

Based on the local and global balance concepts of Zanella (2020), we thus introduced a generalized version of MALA. The new sampler features two tuning parameters: the usual step size σ^2 and an interpolation parameter $\gamma \in [1, 2]$ that accommodates the dimension of the target distribution. The extra parameter adds to the flexibility of the usual MALA by customizing the impact of the biasing term in the proposal distribution, which leads to a computationally-free improvement. We obtained theoretical and empirical results about the tuning of these parameters, which were then used to provide specific and user-friendly tuning recommendations.

Although the traditional MALA is known to be optimal in infinite-dimensional settings, in practice, the annealed MALA remains the most appealing option (even for high-dimensional targets). Numerical illustrations indeed corroborate the existence of relatively consistent efficiency gains; such gains most often range from 10% to 25% compared to MALA, but can also approach 100% in some cases. The efficiency of the annealed MALA also compares favourably to that of MALA in various Bayesian logistic regression contexts. One however needs to be careful and avoid overestimating γ too aggressively as the benefit from implementing the annealed sampler could become marginal, or even negative.

The proposed algorithm does not only generalize the isotropic MALA, but also its more general position-dependent version. This means that where a conditioning matrix is required for efficiently exploring the space, we still benefit from including an interpolation parameter γ. As a general and efficient tuning approach, we suggest adjusting the interpolation parameter as a function of d using $\gamma_d^* = 1 + d^{-\min\{\sqrt{d}/10,1/3\}}$; this guideline has consistently provided near-optimal processes in the examples considered. The step size may then be tuned so as to obtain an acceptance rate that lies in the $40\% - 60\%$ range; unless γ is grossly overestimated, the performance appears to be quite robust to choices of step size in the proposed range. Highly complex targets sometimes require more conservative steps, so we suggest opting for a smaller proposal variance (higher acceptance rate) if necessary. A poor performance under these settings might be indicative of the need for a conditioned version of the sampler in isotropic cases, or simply a better conditioning process when a general MALA is already implemented. We did not encounter any situation where the proposed tunings offered performances significantly worse than MALA, although equivalent performances might arise in cases where potential efficiency gains are modest.

All things considered, the main appeal of the new sampler can be summarized as its low-effort implementation, combined to the fact that it rarely produces an output that is worse than MALA. Using the guidelines presented in this article, one could naturally turn to adaptation to appropriately tune γ and σ^2. In this automated framework, we could even consider a position-dependent parameter $\gamma(x_t)$, or even interpolation parameters $\gamma_1, \ldots, \gamma_d$ adjusted to individual target components; this shall be perused separately. One interpretation of the algorithm introduced is that it builds a proposal based on an annealed version of the target; this appears like an easy-to-generalise perspective and opens interesting directions for future research.

Declarations

Appendix: Proofs of the various results discussed in the article (.pdf file).

Data sets: Data sets used in §4 are publicly available at `archive.ics.uci.edu` (German Credit, Australian Credit, Heart) and in the package `MASS` on `R` (Pima Indian).

Code: `R` code will be made available, TBA.

Declaration of interest: none.

References

Boisvert-Beaudry, G. (2019). Efficacit des distributions instrumentales en quilibre dans un algorithme de type Metropolis-Hastings. Thesis, Université de Montréal.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, 473–483.

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

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

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.

Michie, D., D. J. Spiegelhalter, C. C. Taylor, and J. Campbell (Eds.) (1994). *Machine Learning, Neural and Statistical Classification*. Upper Saddle River, NJ, USA: Ellis Horwood.

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

Ripley, B. D. and N. Hjort (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

Roberts, G. and J. Rosenthal (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science 16*(4), 351–367.

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

Tierney, L. (1998). A note on Metropolis-Hastings kernels for general state spaces. *Annals of Applied Probability*, 1–9.

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

A Appendix: Proofs of Section 2

A.1 Globally-balanced proposal distributions

The following result characterizes the functions g that satisfy the global balance condition. The proof for the locally-balanced case can be found in Zanella (2020).

Theorem A.1. *Let $Q_\sigma(x,\cdot)$ be a d-dimensional symmetrical proposal distribution centered at $x \in \mathcal{S}$, with scale parameter σ. Assume that $q_\sigma(x,\cdot)$, its associated bounded density, can be expressed as*

$$q_\sigma(x,y) = \frac{1}{\sigma^d} r\left(\frac{y-x}{\sigma}\right) \ , \tag{A.1}$$

where r is a unit-scale density such that $r(-z) = r(z)$. Let $g : \mathcal{S} \times \mathcal{S} \to [0,\infty]$ be a bounded, continuous function such that $\int_{\mathcal{S}} g(x,z)\mathrm{d}z < \infty$ for all $x \in \mathcal{S}$. The biased proposal distribution $Q_{g,\sigma}$ in (1) *satisfies the global balance condition if, and only if,*

$$g(x,y) \propto \pi(y), \quad \forall x,y \in \mathcal{S} \ .$$

Proof. ($\Leftarrow$) Suppose $g(x,y) \propto \pi(y)$. Using (1) and the symmetry of q_σ, we have

$$\frac{Z_\sigma(x) q_{g,\sigma}(x,y)\mathrm{d}y\mathrm{d}x}{g(x,y)} = q_\sigma(x,y)\mathrm{d}y\mathrm{d}x = q_\sigma(y,x)\mathrm{d}x\mathrm{d}y = \frac{Z_\sigma(y) q_{g,\sigma}(y,x)\mathrm{d}x\mathrm{d}y}{g(y,x)} \ ,$$

where Z_σ is a normalization constant as detailed in Section 2. The function $g(x,y)$ being proportional to $\pi(y)$ implies that

$$\pi(x) Z_\sigma(x) q_{g,\sigma}(x,y)\mathrm{d}y\mathrm{d}x = \pi(y) Z_\sigma(y) q_{g,\sigma}(y,x)\mathrm{d}x\mathrm{d}y \ ,$$

and the biased proposal distribution thus generates a Markov process that is reversible with respect to the density

$$\frac{\pi(x) Z_\sigma(x)\mathrm{d}x}{\int_{\mathcal{S}} \pi(z) Z_\sigma(z)\mathrm{d}z} = \frac{\sigma^d \pi(x) Z_\sigma(x)\mathrm{d}x}{\sigma^d \int_{\mathcal{S}} \pi(z) Z_\sigma(z)\mathrm{d}z} \ . \tag{A.2}$$

Our interest is in the limit of this distribution as $\sigma \uparrow \infty$. Using the assumption on q_σ, we can rewrite the numerator in (A.2) as

$$\sigma^d \pi(x) Z_\sigma(x) = \pi(x) \int_{\mathcal{S}} g(x,z) \sigma^d q_\sigma(x,z)\mathrm{d}z = \pi(x) \int_{\mathcal{S}} g(x,z) r\left(\frac{z-x}{\sigma}\right)\mathrm{d}z \ .$$

Since $g(x,z) \propto \pi(z)$ and the density r is bounded, we use the Dominated Convergence Theorem to obtain the limit of the last integral; for a constant $0 < c < \infty$,

$$\int_{\mathcal{S}} g(x,z) r\left(\frac{z-x}{\sigma}\right)\mathrm{d}z = \int_{\mathcal{S}} c\pi(z) r\left(\frac{z-x}{\sigma}\right)\mathrm{d}z \xrightarrow{\sigma\uparrow\infty} \int_{\mathcal{S}} c\pi(z) r(0)\mathrm{d}z = cr(0) \ .$$

The limit of the numerator is thus $\pi(x)cr(0)$. We again use the Dominated Convergence Theorem to evaluate the limit of the integral at the denominator of (A.2). Since the density r is bounded by a constant $M < \infty$ (say), the integrand is bounded by

$$\sigma^d \pi(z) Z_\sigma(z) = \pi(z) \int_{\mathcal{S}} c\pi(u) r\left(\frac{u-z}{\sigma}\right) \mathrm{d}u \leqslant cM\pi(z) \ ,$$

which is integrable. Applying the aforementioned theorem and using the limit of $\sigma^d \pi(x) Z_\sigma(x)$ calculated earlier, the limit of the denominator is

$$\lim_{\sigma \uparrow \infty} \sigma^d \int_{\mathcal{S}} \pi(z) Z_\sigma(z) \mathrm{d}z = \int_{\mathcal{S}} \pi(z) cr(0) \mathrm{d}z = cr(0) \ .$$

Combining both limits, it follows that the density associated to (A.2) converges to $\pi(x)$. Hence, by Scheffé's Lemma, the distribution converges to $\Pi(\mathrm{d}x)$.

($\Rightarrow$) Suppose $Q_{g,\sigma}$ is globally balanced with respect to Π. As before, we can write

$$\sigma^d g(y,x) Z_\sigma(x) q_{g,\sigma}(x,y) \mathrm{d}y \mathrm{d}x = \sigma^d g(x,y) Z_\sigma(y) q_{g,\sigma}(y,x) \mathrm{d}x \mathrm{d}y \ , \tag{A.3}$$

including σ^d on both sides of the equation to help with further calculations. By taking the limit as $\sigma \uparrow \infty$ on both sides of (A.3), we will be able to capitalize on the global balance hypothesis. We again use the Dominated Convergence Theorem on the term $\sigma^d Z_\sigma(x)$, since

$$\sigma^d Z_\sigma(x) = \int_{\mathcal{S}} g(x,z) r\left(\frac{z-x}{\sigma}\right) \mathrm{d}z \leqslant M \int_{\mathcal{S}} g(x,z) \mathrm{d}z < \infty \ ,$$

by assumption on g and using the fact that r is bounded. Hence, taking the limit on both sides of (A.3) and noting q_g^* as the asymptotic density of $q_{g,\sigma}$, we get

$$\left(g(y,x) \int_{\mathcal{S}} g(x,z) r(0) \mathrm{d}z \right) q_g^*(x,y) \mathrm{d}y \mathrm{d}x = \left(g(x,y) \int_{\mathcal{S}} g(y,z) r(0) \mathrm{d}z \right) q_g^*(y,x) \mathrm{d}x \mathrm{d}y \ .$$

The term $\int_{\mathcal{S}} g(x,z) r(0) \mathrm{d}z = k(x)$ is a function of x only. Now, being in the asymptotic realm, the global balance hypothesis implies, for a constant $\nu > 0$, that

$$g(y,x) k(x) = \nu \pi(x) \Rightarrow g(y,x) = \frac{\nu \pi(x)}{k(x)} \ ,$$

where $k(x) > 0$ since in the opposite case, the global balance would not be respected. Hence, $g(y,x)$ is a function of x only, which means that $k(x) = \int_{\mathcal{S}} g(x,z) \mathrm{d}z$ is a constant. We conclude that $g(y,x) \propto \pi(x)$ for all $x, y \in \mathcal{S}$. □

A.2 Efficiency of globally-balanced proposal distributions

In order to simplify the expression for the expected squared jumping distance (ESJD) in finite-dimensional contexts, we need to partition the state space $\mathcal{S}$. In particular, because of the specific form of the acceptance probability $\alpha(x, y)$, there are only four distinct regions of $\mathcal{S}$ in which the candidate (y) for the next state of the process can lie:

- **the identity region** $R_{id}(x) := \{x\}$, which is a set of measure zero under $Q_\sigma(x, \cdot)$;
- **the equality region** $R_{eq,\sigma}(x) := \left\{y \in \mathcal{S} : y \notin R_{id}(x), \frac{\pi(y)q_\sigma(y,x)}{\pi(x)q_\sigma(x,y)} = 1\right\}$;
- **the acceptance region** $R_{a,\sigma}(x) := \left\{y \in \mathcal{S} : \frac{\pi(y)q_\sigma(y,x)}{\pi(x)q_\sigma(x,y)} > 1\right\}$, which is the last of the three regions where the candidate is automatically accepted;
- **the rejection region** $R_{r,\sigma}(x) := \{y \in \mathcal{S} : \alpha(x, y) < 1\}$, which is the region where the candidate is not automatically accepted.

Using these regions, we can now prove two lemmas that will then be used in the proof of Proposition 1. The first one, Lemma A.2, mixes results from Lemma 1 and Corollary 1 in Sherlock (2006).

Lemma A.2. *Consider a target density π on $\mathcal{S}$ and suppose that a Metropolis-Hastings with proposal density $q_\sigma(x, \cdot)$ is used to sample from this target. Let X be the current state of the process and Y be the candidate for the next state. Then, under stationarity, the following equality holds*

$$\int_{x\in\mathcal{S}} \int_{y\in R_{a,\sigma}(x)} \|y - x\|^2 A(\mathrm{d}x, \mathrm{d}y) = \int_{x\in\mathcal{S}} \int_{y\in R_{r,\sigma}(x)} \|y - x\|^2 A(\mathrm{d}x, \mathrm{d}y) ,$$

where $A(\mathrm{d}x, \mathrm{d}y) = \pi(x)q_\sigma(x, y)\alpha(x, y)\mathrm{d}y\mathrm{d}x$ denotes the joint density of X and Y. The expected squared jumping distance (ESJD) is then expressed as

$$ESJD = \int_{\mathcal{S}} \left[\int_{R_{eq,\sigma}(x)} \|y - x\|^2 q_\sigma(x, y)\mathrm{d}y + 2 \int_{R_{a,\sigma}(x)} \|y - x\|^2 q_\sigma(x, y)\mathrm{d}y \right] \pi(x)\mathrm{d}x \ .$$

Proof. The first part follows from the reversibility of the Markov process, as well as from the interchangeability of the regions $R_{a,\sigma}(\cdot)$ and $R_{r,\sigma}(\cdot)$. Indeed, we can write

$$\begin{aligned}
\int_{x\in\mathcal{S}} \int_{y\in R_{a,\sigma}(x)} \|y - x\|^2 A(\mathrm{d}x, \mathrm{d}y) &= \int_{x\in\mathcal{S}} \int_{y\in R_{a,\sigma}(x)} \|y - x\|^2 \pi(x)q_\sigma(x, y)\alpha(x, y)\mathrm{d}y\mathrm{d}x \\
&= \int_{x\in\mathcal{S}} \int_{y\in R_{a,\sigma}(x)} \|y - x\|^2 \pi(y)q_\sigma(y, x)\alpha(y, x)\mathrm{d}y\mathrm{d}x \\
&= \int_{y\in\mathcal{S}} \int_{x\in R_{r,\sigma}(y)} \|y - x\|^2 \pi(y)q_\sigma(y, x)\alpha(y, x)\mathrm{d}x\mathrm{d}y \\
&= \int_{y\in\mathcal{S}} \int_{x\in R_{r,\sigma}(y)} \|y - x\|^2 A(\mathrm{d}y, \mathrm{d}x) \ ,
\end{aligned}$$

and then swapping the integration variables on the last line simply leads to

$$\int_{x\in\mathcal{S}}\int_{y\in R_{a,\sigma}(x)} \|y-x\|^2 A(\mathrm{d}x,\mathrm{d}y) = \int_{x\in\mathcal{S}}\int_{y\in R_{r,\sigma}(x)} \|y-x\|^2 A(\mathrm{d}x,\mathrm{d}y) \ .$$

Ignoring the region $R_{id}(x)$ (since $\|y-x\|^2 = 0$ for $y \in R_{id}(x)$), we can write

$$\begin{aligned}
ESJD &= \iint \|y-x\|^2 A(\mathrm{d}x,\mathrm{d}y) \\
&= \int_{x\in\mathcal{S}} \Bigg[\int_{y\in R_{eq,\sigma}(x)} \|y-x\|^2 A(\mathrm{d}x,\mathrm{d}y) + \int_{y\in R_{a,\sigma}(x)} \|y-x\|^2 A(\mathrm{d}x,\mathrm{d}y) \\
&\qquad + \int_{y\in R_{r,\sigma}(x)} \|y-x\|^2 A(\mathrm{d}x,\mathrm{d}y) \Bigg] \\
&= \int_{x\in\mathcal{S}} \Bigg[\int_{y\in R_{eq,\sigma}(x)} \|y-x\|^2 A(\mathrm{d}x,\mathrm{d}y) + 2\int_{y\in R_{a,\sigma}(x)} \|y-x\|^2 A(\mathrm{d}x,\mathrm{d}y) \Bigg] \\
&= \int_{x\in\mathcal{S}} \Bigg[\int_{y\in R_{eq,\sigma}(x)} \|y-x\|^2 q_\sigma(x,y)\mathrm{d}y + 2\int_{y\in R_{a,\sigma}(x)} \|y-x\|^2 q_\sigma(x,y)\mathrm{d}y \Bigg] \pi(x)\mathrm{d}x \ ,
\end{aligned}$$

where we use the fact that for all $x \in \mathcal{S}$, $\alpha(x,y) = 1$ if $y \in \{R_{eq,\sigma}(x) \cup R_{a,\sigma}(x)\}$. □

The regions $R_{a,\sigma}(x)$ and $R_{eq,\sigma}(x)$ are functions of the parameter σ through the proposal density. In order to determine the limiting ESJD when $\sigma^2 \uparrow \infty$, we have to find the limiting form of these regions. This is studied in Lemma A.3 for globally-balanced proposal distributions.

Lemma A.3. *Consider a bounded target density π on $\mathcal{S}$. Suppose that a Metropolis-Hastings algorithm with biased proposal kernel $Q_{g,\sigma}$ is used to sample from this target, where $g(x,y) = \pi(y)$. Furthermore, let*

$$A_\sigma(x) := \{R_{eq,\sigma}(x) \cup R_{a,\sigma}(x)\} = \left\{ y \in \mathcal{S}\backslash\{x\} : \frac{\pi(y)q_{g,\sigma}(y,x)}{\pi(x)q_{g,\sigma}(x,y)} \geqslant 1 \right\} \ . \tag{A.4}$$

Then, for all $x \in \mathcal{S}$, the limit of this set when $\sigma \uparrow \infty$ is

$$\lim_{\sigma\uparrow\infty} A_\sigma(x) = \mathcal{S}\backslash\{x\} \ .$$

Proof. From (1), the biased proposal density satisfies

$$q_{g,\sigma}(x,y) = \frac{\pi(y)q_\sigma(x,y)}{Z_\sigma(x)} \ ,$$

with $Z_\sigma(x) = \int_{\mathcal{S}} \pi(z) q_\sigma(x,z)\mathrm{d}z$. For all $x \in \mathcal{S}$, $q_\sigma(x,\cdot)$ is a symmetrical density that can be expressed as (A.1). Given the form of the regions $R_{eq,\sigma}(x)$ and $R_{a,\sigma}(x)$, we are interested in the ratio

$$\frac{\pi(y) q_{g,\sigma}(y,x)}{\pi(x) q_{g,\sigma}(x,y)} = \frac{\pi(y)\pi(x) q_\sigma(y,x) Z_\sigma(x)}{\pi(x)\pi(y) q_\sigma(x,y) Z_\sigma(y)} = \frac{Z_\sigma(x)}{Z_\sigma(y)} \ . \tag{A.5}$$

We want to obtain the limit of (A.5) when $\sigma \uparrow \infty$. We first write

$$\sigma^d Z_\sigma(x) = \int_{\mathcal{S}} \pi(z) \sigma^d q_\sigma(x,z)\mathrm{d}z = \int_{\mathcal{S}} \pi(z) r\left(\frac{z-x}{\sigma}\right)\mathrm{d}z \xrightarrow{\sigma\uparrow\infty} \int_{\mathcal{S}} \pi(z) r(0)\mathrm{d}z = r(0) \ ,$$

where the convergence follows the same arguments as in §A.1. The limit of (A.5) then becomes

$$\frac{\sigma^d Z_\sigma(x)}{\sigma^d Z_\sigma(y)} \xrightarrow{\sigma\uparrow\infty} 1, \quad \forall x,y \in \mathcal{S} \ .$$

This implies that $\lim_{\sigma\uparrow\infty} A_\sigma(x) = \mathcal{S}\backslash\{x\}$. □

Using the previous lemmas, we can now prove Proposition 1.

Proposition A.4. *Let π be a bounded target density such that $\mathbb{E}_\pi[\|X\|^2] < \infty$, where $\|\cdot\|$ denotes the Euclidean norm. Suppose that a Metropolis-Hastings algorithm with a globally-balanced proposal distribution is used to sample from this target. Let the blinded portion of the proposal density, q_σ, be such that*

$$\pi(y) > 0 \quad \Rightarrow \quad q_\sigma(x,y) > 0 \ , \quad \forall x \in \mathcal{S} \ . \tag{A.6}$$

Then,

$$\lim_{\sigma\uparrow\infty} \mathbb{E}[\|X_{t+1} - X_t\|^2] \quad = \quad \lim_{\sigma\uparrow\infty} \iint \|y-x\|^2 q_{g,\sigma}(x,y)\alpha(x,y)\mathrm{d}y\mathrm{d}x > 0 \ ,$$

where $q_{g,\sigma}$ is the density associated to $Q_{g,\sigma}$.

Proof. Without loss of generality, we suppose that $\sigma \geqslant c$ for some constant $0 < c < \infty$. We wish to find a lower bound that is strictly greater than 0 for the limiting expected squared jumping distance. Using Lemma A.2, we write

$$\begin{aligned} ESJD &= \mathbb{E}[\|Y-X\|^2] \\ &= \int_{x\in\mathcal{S}} \pi(x) \left[\int_{y\in R_{eq,\sigma}(x)} \|y-x\|^2 q_{g,\sigma}(x,y)\mathrm{d}y + 2\int_{y\in R_{a,\sigma}(x)} \|y-x\|^2 q_{g,\sigma}(x,y)\mathrm{d}y \right]\mathrm{d}x \\ &\geqslant \int_{x\in\mathcal{S}} \pi(x) \int_{y\in A_\sigma(x)} \|y-x\|^2 q_{g,\sigma}(x,y)\mathrm{d}y\mathrm{d}x \ , \end{aligned}$$

where the set $A_\sigma(x)$ is as in (A.4). To avoid integration issues, we only consider points $x \in B(0,a)$, the ball of radius $0 < a < \infty$ centered at $(0,\ldots,0)^\top \in \mathcal{S}$. Then,

$$\begin{aligned} ESJD &\geqslant \int_{x\in B(0,a)} \frac{\pi(x)}{Z_\sigma(x)} \int_{y\in A_\sigma(x)} \|y-x\|^2 \pi(y) q_\sigma(x,y) \mathrm{d}y\mathrm{d}x \\ &= \int_{x\in B(0,a)} \frac{\pi(x)}{Z^*_\sigma(x)} \int_{y\in A_\sigma(x)} \|y-x\|^2 \pi(y) r\left(\frac{y-x}{\sigma}\right) \mathrm{d}y\mathrm{d}x \ , \end{aligned} \tag{A.7}$$

where $Z^*_\sigma(x) = \int_{\mathcal{S}} \pi(z)\sigma^d q_\sigma(x,z)\mathrm{d}z$.

To compute the limit of the ESJD as $\sigma \uparrow \infty$, we need to move the limit inside the integrals. For the first integral, we use the Dominated Convergence Theorem; we thus need to find a function (independent of σ) that acts as an upper bound for the function

$$F_\sigma(x) = \frac{\pi(x)}{Z^*_\sigma(x)} \int_{y\in A_\sigma(x)} \|y-x\|^2 \pi(y) r\left(\frac{y-x}{\sigma}\right) \mathrm{d}y \ , \tag{A.8}$$

for all $\sigma \geqslant c$. We first focus on the numerator of (A.8). Since $\sup_{z\in\mathcal{S}} r(z) = M < \infty$, then the numerator of $F_\sigma(x)$ is bounded by

$$\begin{aligned} \pi(x) \int_{y\in A_\sigma(x)} \|y-x\|^2 \pi(y) r\left(\frac{y-x}{\sigma}\right) \mathrm{d}y &\leqslant M\pi(x) \int_{y\in A_\sigma(x)} \|y-x\|^2 \pi(y)\mathrm{d}y \\ &\leqslant M\pi(x) \int_{y\in\mathcal{S}} (\|y\|^2 + \|x\|^2 + 2\|y\|\|x\|)\pi(y)\mathrm{d}y. \end{aligned}$$

Since $M_2 \equiv \mathbb{E}_\pi[\|X\|^2] < \infty$, then by Jensen's inequality $M_1 \equiv \mathbb{E}_\pi[\|X\|] < \infty$; therefore,

$$M\pi(x) \int_{y\in\mathcal{S}} (\|y\|^2 + \|x\|^2 + 2\|y\|\|x\|)\pi(y)\mathrm{d}y \leqslant M\pi(x)(M_2 + \|x\|^2 + 2M_1\|x\|) \ ,$$

and the term on the right-hand side is also integrable.

We now show that the denominator in (A.7), that is the function $Z^*_\sigma(x)$, has a constant lower bound for all $\sigma \geqslant c$. Define the set $C(t) = \{x \in \mathcal{S} : \pi(x) \geqslant t\}$, where $t > 0$ is such that $C(t)$ is of positive Lebesgue measure. Then, we have

$$Z^*_\sigma(x) = \int_{\mathcal{S}} \pi(z)\sigma^d q_\sigma(x,z)\mathrm{d}z \geqslant \int_{C(t)} \pi(z) r\left(\frac{z-x}{\sigma}\right) \mathrm{d}z \geqslant t \int_{C(t)} r\left(\frac{z-x}{\sigma}\right) \mathrm{d}z.$$

Naturally, $C(t)$ is of finite measure as the density π could not integrate to 1 otherwise; indeed, we have

$$1 = \int_{\mathcal{S}} \pi(z)\mathrm{d}z = \int_{C(t)} \pi(z)\mathrm{d}z + \int_{\mathcal{S}\setminus C(t)} \pi(z)\mathrm{d}z \geqslant t \int_{C(t)} \mathrm{d}z \ .$$

Since the density r is bounded above, the Bounded Convergence Theorem implies that

$$\int_{C(t)} r\left(\frac{z-x}{\sigma}\right) \mathrm{d}z \xrightarrow{\sigma\uparrow\infty} r(0) \int_{C(t)} \mathrm{d}z \ ,$$

where $\lim\limits_{\sigma\uparrow\infty} r((z-x)/\sigma) = r(0)$ for all $x \in B(0,a)$ since $a < \infty$. For $0 < \varepsilon < r(0)$, there thus exists σ_0 such that $|r((z-x)/\sigma) - r(0)| < \varepsilon$ for all $\sigma \geqslant \sigma_0$. Hence, for $\sigma \geqslant \sigma_0$,

$$Z^*_\sigma(x) \geqslant t \int_{C(t)} r\left(\frac{z-x}{\sigma}\right) \mathrm{d}z \geqslant t(r(0)-\varepsilon) \int_{C(t)} \mathrm{d}z > 0 \ .$$

When $c \leqslant \sigma < \sigma_0$, then given (A.6) we necessarily have

$$Z^*_\sigma(x) = \int_{\mathcal{S}} \pi(z)\sigma^d q_\sigma(x,z)\mathrm{d}z \geqslant c^d \int_{\mathcal{S}} \pi(z) q_\sigma(x,z)\mathrm{d}z > 0 \ .$$

This means that $m \equiv \inf\limits_{c\leqslant\sigma<\sigma_0} Z^*_\sigma(x) > 0$. Then, for all $\sigma \geqslant c$, we have $Z^*_\sigma(x) \geqslant \min\{m, t(r(0)-\varepsilon)\int_{C(t)} \mathrm{d}z\}$. The function $F_\sigma(x)$ in (A.8) is thus bounded by

$$F_\sigma(x) \leqslant \frac{M\pi(x)(M_2 + \|x\|^2 + 2M_1\|x\|)}{\min\{m, t(r(0)-\varepsilon)\int_{C(t)} \mathrm{d}z\}} \ ,$$

which is independent of σ and integrable on $B(0,a)$.

Using the Dominated Convergence Theorem, we can thus move, in (A.7), the limit inside the first integral

$$\lim_{\sigma\uparrow\infty} ESJD \geqslant \int_{B(0,a)} \lim_{\sigma\uparrow\infty} \frac{\pi(x)}{Z^*_\sigma(x)} \int_{y\in A_\sigma(x)} \|y-x\|^2 \pi(y) r\left(\frac{y-x}{\sigma}\right) \mathrm{d}y\mathrm{d}x \ . \tag{A.9}$$

We can now separately compute the limits of $Z^*_\sigma(x)$ and of the numerator in (A.9). As seen in §A.1, we have

$$Z^*_\sigma(x) = \int_{\mathcal{S}} \pi(z) r\left(\frac{z-x}{\sigma}\right) \mathrm{d}z \xrightarrow{\sigma\uparrow\infty} \int_{\mathcal{S}} \pi(z) r(0) \mathrm{d}z = r(0) \ .$$

To move the limit inside the second integral in (A.9), we use the Dominated Convergence Theorem. We have

$$\begin{aligned} \int_{y\in A_\sigma(x)} \|y-x\|^2 \pi(y) r\left(\frac{y-x}{\sigma}\right) \mathrm{d}y &= \int_{y\in\mathcal{S}} \|y-x\|^2 \pi(y) \mathbb{1}_{y\in A_\sigma(x)} r\left(\frac{y-x}{\sigma}\right) \mathrm{d}y \qquad \text{(A.10)} \\ &\leqslant M \int_{y\in\mathcal{S}} \|y-x\|^2 \pi(y) \mathrm{d}y \ , \end{aligned}$$

where $\|y-x\|^2 \pi(y)$ is independent of σ and integrable. From Lemma A.3, the limit of the set $A_\sigma(x)$ when $\sigma \uparrow \infty$ is $\mathcal{S}\backslash\{x\}$, and so the limit of (A.10) is

$$\int_{y\in A_\sigma(x)} \|y-x\|^2 \pi(y) r\left(\frac{y-x}{\sigma}\right) \mathrm{d}y \xrightarrow{\sigma\uparrow\infty} r(0) \int_{y\in\mathcal{S}\backslash\{x\}} \|y-x\|^2 \pi(y)\mathrm{d}y \ .$$

Combining the limits of $Z^*_\sigma(x)$ and of (A.10), we conclude that

$$\lim_{\sigma\uparrow\infty} ESJD \geqslant \int_{B(0,a)} \pi(x) \int_{y\in\mathcal{S}\backslash\{x\}} \|y-x\|^2 \pi(y)\mathrm{d}y\mathrm{d}x > 0 \ .$$

□

A.3 Approximated kernels: local and global balances

The following result shows how two different choices of the biasing function g lead to the MALA sampler.

Proposition A.5. *The locally-balanced proposal kernels $Q_{g,\sigma}$ obtained by combining the normal kernel $\mathcal{N}(x,\sigma^2 I_d)$ and one of the balancing functions*

$$g_1(x,y) = \sqrt{\frac{\pi(y)}{\pi(x)}} \qquad or \qquad g_2(x,y) = \frac{\pi(y)}{\pi(y)+\pi(x)} \ ,$$

can be approximated by the MALA proposal kernel.

Proof. Consider the case of g_1 first. Since $Q_\sigma(x,\cdot)$ is the normal distribution, we can write

$$\begin{aligned} q_{g_1,\sigma}(x,y)\mathrm{d}y &\propto \left\{\frac{\pi(y)}{\pi(x)}\right\}^{1/2} q_\sigma(x,y)\mathrm{d}y \\ &\propto \exp\left\{\frac{1}{2}\log\{\pi(y)\}\right\} \exp\left\{-\frac{1}{2\sigma^2}(y-x)^\top(y-x)\right\}\mathrm{d}y \ . \qquad \text{(A.11)} \end{aligned}$$

A first-order Taylor approximation of $\log\{\pi(y)\}$ around x gives

$$\pi(y) = \exp\{\log(\pi(y))\} \approx \exp\{\log\{\pi(x)\} + \nabla\log\{\pi(x)\}(y-x)\} \ .$$

This expression is then used to approximate (A.11) as

$$\begin{aligned} q_{g_1,\sigma}(x,y)\mathrm{d}y &\propto \exp\left\{\frac{1}{2}\log\{\pi(y)\}\right\} \exp\left\{-\frac{1}{2\sigma^2}(y-x)^\top(y-x)\right\}\mathrm{d}y \\ &\approx \exp\left\{\frac{1}{2}\log\{\pi(x)\} + \frac{1}{2}\nabla\log\{\pi(x)\}(y-x)\right\} \exp\left\{-\frac{1}{2\sigma^2}(y-x)^\top(y-x)\right\}\mathrm{d}y \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(y-x)^\top(y-x) + \frac{1}{2}\nabla\log\{\pi(x)\}(y-x)\right\}\mathrm{d}y \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\left\|y-x-\frac{\sigma^2}{2}\nabla\log\{\pi(x)\}\right\|^2\right\}\mathrm{d}y \ , \end{aligned}$$

which is proportional to the MALA proposal density.

Now consider the balancing function g_2. Letting $k(y) = \pi(x) + \pi(y)$, the first-order Taylor approximation of g_2 with respect to y around x becomes

$$\begin{aligned} g_2(x,y) &= \exp\left\{ \log\{\pi(y)\} - \log\{\pi(x)+\pi(y)\} \right\} \\ &\approx \exp\left\{ \log\{\pi(x)\} + \nabla \log\{\pi(x)\}(y-x) - \log\{2\pi(x)\} - \nabla(\log\circ k)(x)(y-x) \right\} \\ &\propto \exp\left\{ \nabla \log\{\pi(x)\}(y-x) - \left(\frac{\nabla\pi(x)}{2\pi(x)}\right)(y-x) \right\} \\ &= \exp\left\{ \frac{1}{2}\nabla \log\{\pi(x)\}(y-x) \right\} \ . \end{aligned}$$

With this approximation for g_2, the locally-balanced proposal density becomes

$$\begin{aligned} q_{g_2,\sigma}(x,y)\mathrm{d}y &\approx \exp\left\{ \frac{1}{2}\nabla \log\{\pi(x)\}(y-x) \right\} \exp\left\{ -\frac{1}{2\sigma^2}(y-x)^\top(y-x) \right\} \mathrm{d}y \\ &\propto \exp\left\{ -\frac{1}{2\sigma^2} \left\| y - x - \frac{\sigma^2}{2}\nabla\log\{\pi(x)\} \right\|^2 \right\} \mathrm{d}y \ , \end{aligned}$$

which is again proportional to the MALA proposal density. □

B Appendix: Proofs of Section 3

B.1 Scaling function of the annealed MALA

To prove Theorem 2 (and subsequently Theorem 4), we need the following lemma whose proof can be found in Beskos and Stuart (2009).

Lemma B.6. *Let $T \in \mathbb{R}$ be a random variable. By defining $x \wedge y := \min\{x, y\}$, we have*

1. *For every $c > 0$,*

$$\mathbb{E}[1 \wedge \mathrm{e}^T] \geqslant \mathrm{e}^{-c}\left(1 - \frac{\mathbb{E}[|T|]}{c}\right) \ .$$

2. *If $\mathbb{E}[T] < 0$, then*

$$\mathbb{E}[1 \wedge \mathrm{e}^T] \leqslant \mathrm{e}^{\mathbb{E}[T]/2} + \frac{2\mathbb{E}[|T - \mathbb{E}[T]|]}{(-\mathbb{E}[T])} \ .$$

Proof of Theorem 2. Let $X = x$ be the current state of the generated Markov process; since the latter is assumed to start in stationarity, this means that $X \sim \Pi$. A candidate y is proposed, and then accepted with probability $\alpha(x, y) = 1 \wedge \mathrm{e}^{R_d}$. The densities π and q_{γ,σ_d} having i.i.d. components, R_d can be written as

$$R_d = \log\left(\frac{\pi(y)q_{\gamma,\sigma_d}(y,x)}{\pi(x)q_{\gamma,\sigma_d}(x,y)}\right) = \sum_{i=1}^{d} \log\left(\frac{f(y_i)q^*_{\gamma,\sigma_d}(y_i,x_i)}{f(x_i)q^*_{\gamma,\sigma_d}(x_i,y_i)}\right) \ ,$$

where q^*_{γ,σ_d} is the unidimensional version of the proposal density in (5). More precisely, the ith component of the candidate Y satisfies

$$Y_i = x_i + \frac{\gamma\sigma_d^2}{2} l'(x_i) + \sigma_d Z_i \ ,$$

with $Z_i \sim \mathcal{N}(0, 1)$ independant of $X_i \sim f,\ i = 1, \ldots, d$ and $\gamma \in (1, 2]$. Considering R_d as a function of σ_d, we now use Taylor expansions to determine exactly for which β values (in $\sigma_d^2 = \ell^2/d^\beta$) does $\mathbb{E}[\alpha(X, Y)]$ becomes null as $d \uparrow \infty$.

Case 1: $\sigma_d^2 \propto d^{-\beta}$ with $\beta \geqslant 1$.

We need to show that $\lim_{d\to\infty} \mathbb{E}[\alpha(X, Y)] > 0$. Following the first part of Lemma B.6, we are certain that this is true whenever $\lim_{d\to\infty} \mathbb{E}[|R_d|] < \infty$. Let us consider a second-order Taylor development of R_d around in $\sigma_d = 0$. The details of this calculation can be found in Appendix A.1 of Boisvert-Beaudry (2019). We have

$$R_d = \mathcal{A}_{1,d} + \mathcal{A}_{2,d} + U_d \ ,$$

with

$$\mathcal{A}_{1,d} = \sigma_d R'_d(0) = \sigma_d \sum_{i=1}^{d} C_{1,i} \ , \qquad \text{where } C_{1,i} = l'(x_i)Z_i(1-\gamma) \ ; \tag{B.1}$$

$$\mathcal{A}_{2,d} = \frac{\sigma_d^2}{2!} R''_d(0) = \frac{\sigma_d^2}{2} \sum_{i=1}^{d} C_{2,i} \ , \qquad \text{where } C_{2,i} = (1-\gamma)[l''(x_i)Z_i^2 + l'(x_i)^2\gamma] \ ; \tag{B.2}$$

$$U_d = \frac{\sigma_d^3}{3!} \sum_{i=1}^{d} U_{i,d}(x_i, Z_i, \sigma_i^*) \ ,$$

where $\sigma_i^* \in [0, \sigma_d]$, $i = 1, \ldots, d$. The moments of $C_{1,i}$ and $C_{2,i}$ are bounded because of the conditions on the moments of f and the derivatives of l. Since $|R_d| \leqslant |\mathcal{A}_{1,d}| + |\mathcal{A}_{2,d}| + |U_d|$, we need to verify that the expectation of each term remains finite in the limit as $d \to \infty$.

Upon examination of the terms $U_{i,d}(x_i, Z_i, \sigma_i^*)$, $i = 1, \ldots, d$, we note that they are polynomials of Z_i, derivatives of l, and positive powers of σ_i^*. By the polynomial bound assumption and following the proof of Theorem 1 in Beskos and Stuart (2009), there thus exists polynomials M_1, M_2, and M_3 such that

$$|U_{i,d}(x_i, Z_i, \sigma_i^*)| \leqslant M_1(x_i)M_2(Z_i)M_3(\sigma_i^*) \ .$$

Since X_i and Z_i are independant and all moments of f are finite, then $\mathbb{E}[M_1(X_i)M_2(Z_i)] = \mathbb{E}[M_1(X_i)]\mathbb{E}[M_2(Z_i)] < \infty$ for all i. Furthermore, since $\beta > 0$, there exists an $\varepsilon > 0$ such that $\sigma_i^* < \sigma_d < \varepsilon$ for all i, so $M_3(\sigma_i^*)$ is also bounded by a constant. Therefore, $\mathbb{E}[|U_{i,d}(X_i, Z_i, \sigma_i^*)|] < K < \infty$ for a constant K independant of i and d. Since $\beta \geqslant 1$, the residual term satisfies

$$\lim_{d\to\infty} \mathbb{E}[|U_d|] \leqslant \lim_{d\to\infty} \frac{\sigma_d^3}{3!} dK = 0 \ .$$

We now study the term $|\mathcal{A}_{1,d}|$. Using Jensen's inequality, we obtain

$$\mathbb{E}[|\mathcal{A}_{1,d}|] = \mathbb{E}\left[\sqrt{\mathcal{A}_{1,d}^2}\right] \leqslant \mathbb{E}[\mathcal{A}_{1,d}^2]^{1/2} = \mathbb{E}\left[\left(\sigma_d \sum_{i=1}^d C_{1,i}\right)^2\right]^{1/2} \ .$$

We note that $C_{1,i}$, $i = 1, \ldots, d$ are i.i.d., since each term depends on X_i and Z_i only; moreover, we observe that $\mathbb{E}[C_{1,i}] = 0$. This thus leads to the simpler bound

$$\mathbb{E}[|\mathcal{A}_{1,d}|] \leqslant \sigma_d \left(\sum_{i=1}^d \mathbb{E}[C_{1,i}^2] + \sum_{i=1}^d \sum_{j\neq i} \mathbb{E}[C_{1,i}C_{1,j}]\right)^{1/2} = \sigma_d \sqrt{d}\mathbb{E}[C_{1,1}^2]^{1/2} \ .$$

Since all moments of $C_{1,1}$ are bounded and $\beta \geqslant 1$, it follows that the limit of this term is also bounded.

Now, using the fact that $C_{2,i}$ $(i = 1, \ldots, d)$ are i.i.d. along with the triangle inequality, and then applying Jensen's inequality, we find the following bound on the expectation of $|\mathcal{A}_{2,d}|$

$$\mathbb{E}[|\mathcal{A}_{2,d}|] \leqslant \frac{d\sigma_d^2}{2}\mathbb{E}[|C_{2,1}|] \leqslant \frac{d\sigma_d^2}{2}\mathbb{E}[C_{2,1}^2]^{1/2} \ ,$$

and the limit of $\mathbb{E}[|\mathcal{A}_{2,d}|]$ is also bounded. We deduce that $\lim_{d\to\infty} \mathbb{E}[|R_d|] < \infty$; by Lemma B.6, we conclude that

$$\lim_{d\to\infty} \mathbb{E}[\alpha(X, Y)] > 0 \ .$$

Case 2: $\sigma_d^2 \propto d^{-\beta}$ with $\beta \in (0, 1)$.

We need to show that $\lim_{d\to\infty} \mathbb{E}[\alpha(X,Y)] = 0$. In this case, using the second part of Lemma B.6, it is sufficient to verify that

$$\mathbb{E}[R_d] \xrightarrow{d\uparrow\infty} -\infty \qquad \text{and} \qquad \frac{\mathbb{E}[|R_d - \mathbb{E}[R_d]|]}{-\mathbb{E}[R_d]} \xrightarrow{d\uparrow\infty} 0 \ . \tag{B.3}$$

We first focus on $\mathbb{E}[R_d]$. Using a Taylor expansion of order m of R_d around $\sigma_d = 0$, where $m \in \mathbb{N}$ is such that $(m+1)\beta > 2$, we find

$$R_d = \sum_{j=1}^{m} \mathcal{A}_{j,d} + U_d^* \ ,$$

with

$$\mathcal{A}_{j,d} = \frac{\sigma_d^j}{j!} R_d^{(j)}(0) = \frac{\sigma_d^j}{j!} \sum_{i=1}^{d} C_{j,i} \ ,$$
$$U_d^* = \frac{\sigma_d^{m+1}}{(m+1)!} R_d^{(m+1)}(\sigma^*) = \frac{\sigma_d^{m+1}}{(m+1)!} \sum_{i=1}^{d} U_{i,d}^*(x_i, Z_i, \sigma_i^*) \ ,$$

and $\sigma_i^* \in [0, \sigma_d]$, $i = 1, \ldots, d$. The terms $C_{1,i}$ and $C_{2,i}$ are identical to (B.1) and (B.2).

We first study the term U_d^*. Using arguments similar to those applied in Case 1, we find that the residual term $\mathbb{E}[|U_{i,d}^*(X_i, Z_i, \sigma_i^*)|]$ can be bounded by a constant K_0 that is independant of d and i. The mth order of the expansion implies then implies that

$$\lim_{d\to\infty} \mathbb{E}[U_d^*] \leqslant \lim_{d\to\infty} \mathbb{E}[|U_d^*|] \leqslant \lim_{d\to\infty} \frac{\sigma^{m+1}}{(m+1)!} d K_0 = 0 \ ,$$

and $\mathbb{E}[U_d^*]$ is $\mathcal{O}\left(d^{1-\frac{(m+1)\beta}{2}}\right)$.

For the other terms, since $\mathbb{E}[C_{1,1}] = 0$, we have

$$\mathbb{E}[R_d] = \sum_{j=1}^{m} \mathbb{E}[\mathcal{A}_{j,d}] + \mathbb{E}[U_d^*] = \sum_{j=2}^{m} \frac{d\sigma_d^j}{j!} \mathbb{E}[C_{j,1}] + \mathbb{E}[U_d^*] \ .$$

As before, the terms $C_{j,1}$, $j = 1, \ldots, m$ are all polynomial functions of Z_1 and the derivatives of l. Thus, similarly as in Case 1, all moments of $C_{j,1}$ are bounded. We can therefore deduce that the dominant term of $\mathbb{E}[R_d]$ is $\mathbb{E}[\mathcal{A}_{2,d}]$, which turns out to be negative. Indeed, using $\mathbb{E}[l''(X)] = -\mathbb{E}[l'(X)^2]$ (see the proof of Lemma 6 in Bédard (2007)) and (B.2), we can write

$$\begin{aligned} \mathbb{E}[C_{2,1}] &= (1-\gamma)\mathbb{E}[l''(X_1)Z_1^2 + l'(X_1)^2\gamma] \\ &= (1-\gamma)\left\{\mathbb{E}[l''(X_1)] + \mathbb{E}[l'(X_1)^2]\gamma\right\} \\ &= -(1-\gamma)^2\mathbb{E}[l'(X_1)^2] < 0 \ . \end{aligned}$$

This implies that $\mathbb{E}[R_d] \to -\infty$ at rate $\mathcal{O}(d^{1-\beta})$.

We now study the ratio $\mathbb{E}[|R_d - \mathbb{E}[R_d]|]/-\mathbb{E}[R_d]$. Making use of the triangle inequality and then of Jensen's inequality, we bound the numerator as follows

$$\begin{aligned}
\mathbb{E}[|R_d - \mathbb{E}[R_d]|] &\leqslant \sum_{j=1}^{m} \mathbb{E}[|\mathcal{A}_{j,d} - \mathbb{E}[\mathcal{A}_{j,d}]|] + \mathbb{E}[|U_d^* - \mathbb{E}[U_d^*]|] \\
&\leqslant \sum_{j=1}^{m} \mathbb{V}(\mathcal{A}_{j,d})^{1/2} + \mathbb{E}[|U_d^* - \mathbb{E}[U_d^*]|] \\
&= \sum_{j=1}^{m} \mathbb{V}\left(\frac{\sigma_d^j}{j!}\sum_{i=1}^{d} C_{j,i}\right)^{1/2} + \mathbb{E}[|U_d^* - \mathbb{E}[U_d^*]|] \\
&= \sum_{j=1}^{m} \frac{\sigma_d^j}{j!}\sqrt{d}\,\mathbb{V}(C_{j,1})^{1/2} + \mathbb{E}[|U_d^* - \mathbb{E}[U_d^*]|] \ . \qquad \text{(B.4)}
\end{aligned}$$

The first term of (B.4) being dominant, the numerator of (B.3) is $\mathcal{O}(d^{\frac{1-\beta}{2}})$ while the denominator is $\mathcal{O}(d^{1-\beta})$. Using the fact that $\beta \in (0,1)$, this means that the ratio converges to 0 according to $\mathcal{O}(d^{\frac{\beta-1}{2}})$. By Lemma B.6, we conclude that

$$\lim_{d\to\infty} \mathbb{E}[\alpha(X,Y)] = 0 \ .$$

Since the smallest value of β that gives a positive asymptotic acceptance rate is 1, we conclude that $\beta_0 = 1$.

□

B.2 Tuning of the annealed MALA

To obtain weak convergence and optimal scaling results for the annealed MALA, we consider target densities as in Section 3.1.

Given the current state $X_t = x$, the annealed MALA generates a candidate $Y_{t+1} = y$ from the proposal distribution

$$Y_{t+1} = X_t + \sigma_d Z_{t+1} + \frac{\gamma\sigma_d^2}{2}\nabla\log\{\pi(X_t)\} \ ,$$

with $Z_{t+1} \sim \mathcal{N}(0, I_d)$. It then accepts the candidate (i.e. $X_{t+1} = Y_{t+1}$) with probability

$$\alpha(X_t, Y_{t+1}) = 1 \wedge \frac{\pi(Y_{t+1})q(Y_{t+1}, X_t)}{\pi(X_t)q(X_t, Y_{t+1})} \ ,$$

where

$$q(x,y) = \prod_{i=1}^{d} q(x_i, y_i) = (2\pi\sigma_d^2)^{-d/2}\exp\left\{-\frac{1}{2\sigma_d^2}\left\|y - x - \frac{\gamma\sigma_d^2}{2}\nabla\log\{\pi(x)\}\right\|^2\right\} \ .$$

To find the asymptotically optimal ℓ in $\sigma_d^2 = \ell^2/d$, we need to study the asymptotic behaviour of the Markov process as $d \uparrow \infty$. To compare the discrete-time process to a limiting continuous-time process, it is convenient to work with a sped up version of the initial algorithm. Let $Z_d(t)$ be the time-t value of the process speeded up by a factor of d; in particular, $Z_d(t) = (X_1([dt]), \ldots, X_d([dt]))$, where $[\cdot]$ is the floor function. Instead of proposing a single move per unit time interval, the accelerated process has the possibility of moving, on average, d times.

Theorem B.7 studies the limiting behaviour of $\{Z_1(t); t \geq 0\}$, the first component of $Z_d(t)$, as $d \uparrow \infty$. Corollary B.8 then transforms the weak convergence result into a statement about the efficiency of the sampler as a function of the acceptance rate, as was done in Roberts and Rosenthal (1998). We denote weak convergence in the Skorokhod topology by $\Rightarrow$, standard Brownian motion at time t by $B(t)$ and the standard normal c.d.f. by $\Phi(\cdot)$.

Theorem B.7. *As $d \uparrow \infty$, the process Z_1 converges weakly, in the Skorokhod topology, to the Langevin diffusion defined by*

$$\mathrm{d}Z(t) = \upsilon(\ell, \gamma)^{1/2} \mathrm{d}B(t) + \frac{1}{2} \upsilon(\ell, \gamma) (\log\{f(Z(t))\})' \mathrm{d}t \ ,$$

where $\upsilon(\ell, \gamma) = 2\ell^2 \Phi(-\frac{\ell}{2}(\gamma - 1)\sqrt{\mathbb{E}[\{(\log f(X))'\}^2]})$ is the speed of the limiting diffusion. Furthermore, $\upsilon(\ell, \gamma)$ is maximized at the unique value $\hat{\ell}_\gamma = 2.38/\{(\gamma-1)\sqrt{\mathbb{E}[\{(\log f(X))'\}^2]}$.

We define the expected acceptance rate of the sampler (under stationarity) as follows

$$a_d(\ell) = \mathbb{E}[\alpha(X, Y)] = \iint \pi(x) q(x, y) \alpha(x, y) \mathrm{d}x \mathrm{d}y \ .$$

Corollary B.8. *We have $\lim_{d \to \infty} a_d(\ell) = a(\ell)$, where*

$$a(\ell) = 2\Phi\left(-\frac{1}{2}\ell(\gamma - 1)\sqrt{\mathbb{E}[\{(\log f(X))'\}^2]}\right) \ ,$$

and the expectation is taken over X having density f. The asymptotically optimal acceptance rate is therefore $a(\hat{\ell}_\gamma) = 0.234$.

The proof of Theorem B.7 and Corollary B.8 is simply a mix of the optimal scaling proofs for RWMH and MALA found in Bédard (2007) and Roberts and Rosenthal (1998). Accordingly, we just outline the main steps and avoid the technical treatment of errors. To learn more about the latter, we refer the reader to the original papers.

The proof is based on the theory exposed in Chapter 4 of Ethier and Kurtz (1986), and in particular Theorem 8.2 and Corollary 8.6. This theory roughly says that for the finite-dimensional distributions of a sequence of processes to converge weakly to those of some Markov process, we simply need to verify the $\mathcal{L}^1$-convergence of their generators. Then, further conditions are verified to make sure that the sequence of processes is relatively compact, which leads to the weak convergence of the stochastic processes themselves.

The main task is then to focus on the $\mathcal{L}^1$-convergence of the generators. Hereafter, we omit time indexing and use boldface to denote d-dimensional vectors. The discrete-time generator of the process produced by the accelerated version of the annealed MALA is expressed as

$$\begin{aligned} Gh(\mathbf{x}) &= d\mathbb{E}\left[(h(Y_1) - h(x_1))\,\alpha(\mathbf{x}, \mathbf{Y})\right] \\ &= d\mathbb{E}\left[(h(Y_1) - h(x_1))\left\{1 \wedge \frac{\pi(\mathbf{Y})q(\mathbf{Y}, \mathbf{x})}{\pi(\mathbf{x})q(\mathbf{x}, \mathbf{Y})}\right\}\right] \;. \end{aligned} \tag{B.5}$$

The generator of a Langevin diffusion process with speed measure $\upsilon(\ell)$ is

$$G_L h(x_1) = \upsilon(\ell, \gamma)\left[\frac{1}{2}h''(x_1) + \frac{1}{2}h'(x_1)(\log f(x_1))'\right] \;.$$

The goal is thus to show that

$$\lim_{d\to\infty} \mathbb{E}\left[|Gh(\mathbf{x}) - G_L h(x_1)|\right] = 0 \;.$$

The generators are written in terms of an arbitrary test function h, which can usually be any smooth function. In our case, since the limiting process obtained in a diffusion, it follows that C_c^∞ is a core for the generator, which means that we can focus on functions in that core to prove the above expression.

The expectation in (B.5) is computed with respect to the d-dimensional candidate random variable $\mathbf{Y} = \mathbf{x} + \sigma\mathbf{Z} + \frac{\gamma\sigma^2}{2}\nabla \log \pi(\mathbf{x})$. By Taylor expanding $\alpha(\mathbf{x}, \mathbf{Y})$ with respect to Y_1 and around x_1, we can show that

$$\lim_{d\to\infty} \mathbb{E}\left[\left|Gh(\mathbf{x}) - \tilde{G}h(\mathbf{x})\right|\right] = 0 \;,$$

where

$$\begin{aligned} \tilde{G}h(\mathbf{x}) &= d\mathbb{E}\left[(h(Y_1) - h(x_1))\left\{\mathbb{E}\left[1 \wedge \mathrm{e}^{\sum_{i=2}^d \xi(x_i, Y_i)}\right] + (Y_1 - x_1)\times\right.\right. \\ &\qquad \left.\left.\left(\tfrac{\partial}{\partial Y_1}\xi(x_1, Y_1)\right)_{Y_1 = x_1} \mathbb{E}\left[\mathrm{e}^{\sum_{i=2}^d \xi(x_i, Y_i)}\mathbb{1}\left(\textstyle\sum_{i=2}^d \xi(x_i, Y_i) < 0\right)\right]\right\}\right] \;, \end{aligned} \tag{B.6}$$

$\xi(x_i, Y_i) = \log \frac{f(Y_i)q(Y_i, x_i)}{f(x_i)q(x_i, Y_i)}$, and $\mathbb{1}(\cdot)$ is the indicator function. Now, by Taylor expanding the function $h(Y_1) - h(x_1)$ with respect to Y_1 around x_1, we obtain

$$h(Y_1) - h(x_1) \approx \frac{\ell}{d^{1/2}}h'(x_1)Z_1 + \frac{\ell^2}{d}\frac{\gamma}{2}(\log f(x_1))'h'(x_1) + \frac{\ell^2}{d}\frac{1}{2}h''(x_1)Z_1^2 \;.$$

Substituting this expression in (B.6), we can show that

$$\lim_{d\to\infty} \mathbb{E}\left[\left|\tilde{G}h(\mathbf{x}) - \hat{G}h(\mathbf{x})\right|\right] = 0 \ ,$$

where

$$\begin{aligned}\hat{G}h(\mathbf{x}) &= \frac{\ell^2}{2}\left\{\gamma h'(x_1)(\log f(x_1))' + h''(x_1)\right\}\mathbb{E}\left[1 \wedge \mathrm{e}^{\sum_{i=2}^{d}\xi(x_i,Y_i)}\right] \\ &\quad + \ell^2 h'(x_1)(1-\gamma)(\log f(x_1))'\mathbb{E}\left[\mathrm{e}^{\sum_{i=2}^{d}\xi(x_i,Y_i)}\mathbb{1}\left(\textstyle\sum_{i=2}^{d}\xi(x_i,Y_i) < 0\right)\right] . \quad \text{(B.7)}\end{aligned}$$

All is left to do is computing the expectations with respect to the random variables $Y_2, \ldots, Y_d$ in the previous expression. Again using a Taylor expansion, we find that

$$\xi(x_i, Y_i) \approx \frac{\ell}{d}(\gamma-1)Z_i(\log f(x_i))' + \frac{\ell^2}{2d}(\gamma-1)\left[\gamma\{(\log f(x_i))'\}^2 + Z_i^2(\log f(x_i))''\right] \ .$$

Using the fact that Z_i, $i = 2, \ldots, d$ are independent standard normal, we find that $\sum_{i=2}^{d}\xi(x_i, Y_i)$ also is normally distributed. It is then easy to show (see Roberts and Rosenthal (1998); Bédard (2007)) that

$$\lim_{d\to\infty}\left|\mathbb{E}\left[1 \wedge \mathrm{e}^{\sum_{i=2}^{d}\xi(x_i,Y_i)}\right] - 2\Phi\left(-\frac{\ell}{2}(\gamma-1)\sqrt{\mathbb{E}\left[\{(\log f(X))'\}^2\right]}\right)\right| = 0 \ ,$$

and

$$\lim_{d\to\infty}\left|\mathbb{E}\left[\mathrm{e}^{\sum_{i=2}^{d}\xi(x_i,Y_i)}\mathbb{1}\left(\textstyle\sum_{i=2}^{d}\xi(x_i,Y_i) < 0\right)\right] - \Phi\left(-\frac{\ell}{2}(\gamma-1)\sqrt{\mathbb{E}\left[\{(\log f(X))'\}^2\right]}\right)\right| = 0 \ .$$

Replacing the expectations in (B.7) by their limits in terms of $\Phi(\cdot)$, we find

$$G_L h(x_1) = 2\ell^2\Phi\left(-\frac{\ell}{2}(\gamma-1)\sqrt{\mathbb{E}\left[\{(\log f(X))'\}^2\right]}\right)\left\{\frac{1}{2}h'(x_1) + \frac{1}{2}(\log f(x_1))'h''(x_1)\right\} \ ,$$

and we can show that $\lim_{d\to\infty}\mathbb{E}\left[\left|\hat{G}h(\mathbf{x} - G_L h(\mathbf{x})\right|\right] = 0$, which concludes the convergence proof of Theorem B.7.

The speed limit of the Langevin diffusion, $\upsilon(\ell,\gamma) = 2\ell^2\Phi\left(-\frac{\ell}{2}(\gamma-1)\sqrt{\mathbb{E}\left[\{(\log f(X))'\}^2\right]}\right)$, is optimized at the unique value $\hat{\ell} = \frac{2.38}{(\gamma-1)\sqrt{\mathbb{E}[\{(\log f(X))'\}^2]}}$.

The asymptotically optimal acceptance rate may be easily computed from there. By using the Taylor expansion of $\xi(x_i, Y_i)$ expressed above and reapplying similar arguments as before (but without worrying about the function h, which simplifies things),

$$\lim_{d\to\infty}|a_d(\ell) - a(\ell)| = \lim_{d\to\infty}\left|\mathbb{E}\left[1 \wedge \mathrm{e}^{\sum_{i=1}^{d}\xi(x_i,Y_i)}\right] - 2\Phi\left(-\frac{\ell}{2}(\gamma-1)\sqrt{\mathbb{E}\left[\{(\log f(X))'\}^2\right]}\right)\right| = 0 \ .$$

The asymptotically optimal acceptance rate is therefore the value $a(\hat{\ell}_\gamma) = 0.234$.

B.3 Convergence rate of the interpolation parameter γ_d

Proof of Theorem 4. As in the proof of Theorem 2, we consider the acceptance probability $\alpha(x, y) = 1 \wedge \mathrm{e}^{R_d}$ of a Markov process currently at x, with candidate step y. We desire to apply Lemma B.6 on a Taylor expansion centered at $\sigma_d = 0$ of

$$R_d = \log\left(\frac{\pi(y) q_{\gamma_d,\sigma_d}(y, x)}{\pi(x) q_{\gamma_d,\sigma_d}(x, y)}\right) = \sum_{i=1}^{d} \log\left(\frac{f(y_i) q^*_{\gamma_d,\sigma_d}(y_i, x_i)}{f(x_i) q^*_{\gamma_d,\sigma_d}(x_i, y_i)}\right) ,$$

where q^* is the unidimensional version of the proposal distribution. This time, the ith component of the candidate Y is given by

$$Y_i = x_i + \frac{\gamma_d \sigma_d^2}{2} l'(x_i) + \sigma_d Z_i \ , \quad i = 1, \ldots, d \ ,$$

with $Z_i \sim \mathcal{N}(0, 1)$ independant of $X_i \sim f$, $i = 1, ..., d$. Since σ_d^2 is $\mathcal{O}(d^{-1/3})$ in this case, an expansion of order 6 is required.

($\Rightarrow$) Following Lemma B.6, we want to show that $\lim_{d\to\infty} \mathbb{E}[|R_d|] < \infty$. The Taylor expansion of order 6 of R_d evaluated at $\sigma_d = 0$ gives

$$R_d = \sum_{i=1}^{6} \mathcal{A}_{i,d} + U_d \ ,$$

where for $i = 1, \ldots, 6$, we have

$$\mathcal{A}_{i,d} = \frac{\sigma_d^i}{i!} R^{(i)}(0) = \frac{\sigma_d^i}{i!} \sum_{j=1}^{d} C_{i,j} \qquad \text{and} \qquad U_d = \frac{\sigma_d^7}{7!} \sum_{j=1}^{d} U_{j,d}(x_j, Z_j, \sigma_j^*) \ ,$$

with $\sigma_j^* \in [0, \sigma_d]$, $j = 1, \ldots, d$. Expressions for $C_{1,j}$ and $C_{2,j}$ may be found in (B.1) and (B.2), while the terms $C_{i,j}$, $i = 3, \ldots, 6$, are detailed in Section A.2 of Boisvert-Beaudry (2019).

From Lemmas A.2.1 and A.2.2 of Boisvert-Beaudry (2019), we also know that $\lim_{d\to\infty} \mathbb{E}[C_{i,1}] < \infty$, $i = 1, \ldots, 6$, with $\mathbb{E}[C_{1,1}] = \mathbb{E}[C_{3,1}] = \mathbb{E}[C_{5,1}] = 0$,

$$\mathbb{E}[C_{2,1}] = -(\gamma_d - 1)^2 \mathbb{E}[l'(X_1)^2] \ , \qquad \text{and} \qquad \mathbb{E}[C_{4,1}] = (\gamma_d - 1) K_{4,d} \ , \tag{B.8}$$

where $\lim_{d\to\infty} K_{4,d} < \infty$. We also have $\lim_{d\to\infty} \mathbb{E}[C_{i,1}^2] < \infty$, $i = 1, \ldots, 6$, with

$$\mathbb{E}[C_{1,1}^2] = (\gamma_d - 1)^2 K_1 \qquad \text{and} \qquad \mathbb{E}[C_{2,1}^2] = (\gamma_d - 1)^2 K_{2,d}, \tag{B.9}$$

where $K_1 < \infty$ is independant of d and $\lim_{d\to\infty} K_{2,d} < \infty$.

With that in mind, we start by looking at the expectation of $|U_d|$. The terms $U_{j,d}(x_j, Z_j, \sigma_j^*)$, $j = 1, \ldots, d$, are polynomial functions of Z_j, derivatives of l and positive powers of σ_j^*. Reusing the argument in the proof of Theorem 2, there exists polynomials M_1, M_2, and M_3 such that

$$|U_{j,d}(x_j, Z_j, \sigma_j^*)| \leqslant M_1(x_j) M_2(Z_j) M_3(\sigma_j^*) \ , \qquad j = 1, \ldots, d \ .$$

Since all moments of f are finite and by the independance of X_j and Z_j, we have $\mathbb{E}[M_1(X_j)M_2(Z_j)] < \infty$ for all j. Furthermore, since $\sigma_j^* \leqslant \sigma_d \leqslant \ell$, this means that $M_3(\sigma_j^*) \leqslant K < \infty$ for a constant K independant of j and d. There thus exists a constant K_0 independant of d such that

$$\lim_{d \to \infty} \mathbb{E}[|U_d|] \leqslant \lim_{d \to \infty} \frac{d\sigma_d^7}{7!} K_0 = 0,$$

since $\sigma_d^2 \propto d^{-1/3}$.

Consider now the terms $\mathcal{A}_{i,d}$, $i = 1, \ldots, 5$. By Jensen's inequality, we have

$$\mathbb{E}[|\mathcal{A}_{i,d}|] \leqslant \mathbb{E}[\mathcal{A}_{i,d}^2]^{1/2} = \frac{\sigma_d^i}{i!} \mathbb{E}\left[\left(\sum_{j=1}^{d} C_{i,j}\right)^2\right]^{1/2} \ . \tag{B.10}$$

Using the fact that $C_{i,j}$ ($j = 1, \ldots, d$) are i.i.d., we bound (B.10) by

$$\begin{aligned} \mathbb{E}[|\mathcal{A}_{i,d}|] &\leqslant \frac{\sigma_d^i}{i!} \left\{ d\mathbb{E}[C_{i,1}^2] + \sum_{j=1}^{d} \sum_{k \neq j} \mathbb{E}[C_{i,j} C_{i,k}] \right\}^{1/2} \\ &\leqslant \frac{\sigma_d^i}{i!} \left\{ d\mathbb{E}[C_{i,1}^2] + d^2 \mathbb{E}[C_{i,1}]^2 \right\}^{1/2} \\ &\leqslant \frac{\sigma_d^i}{i!} \left\{ \sqrt{d} \mathbb{E}[C_{i,1}^2]^{1/2} + d|\mathbb{E}[C_{i,1}]| \right\} \ . \end{aligned} \tag{B.11}$$

For $i = 1, 3, 5$, the first moment of $C_{i,1}$ is null and so (B.11) can be bounded by

$$\mathbb{E}[|A_{i,d}|] \leqslant \frac{\sqrt{d}\sigma_d^i}{i!} \mathbb{E}[C_{i,1}^2]^{1/2} = d^{\frac{1}{2} - \frac{i}{6}} \frac{\ell^{\frac{i}{2}}}{i!} \mathbb{E}[C_{i,1}^2]^{1/2} \ .$$

Since $\lim_{d \to \infty} \mathbb{E}[C_{i,1}^2] < \infty$ for $i = 1, \ldots, 6$, we have directly that $\lim_{d \to \infty} \mathbb{E}[|\mathcal{A}_{i,d}|] < \infty$ for $i = 3, 5$. For $i = 1$, using (B.9), we get the following bound

$$\mathbb{E}[|\mathcal{A}_{1,d}|] \leqslant \sqrt{d}\sigma_d \mathbb{E}[C_{1,1}^2]^{1/2} = \sqrt{d}\sigma_d(\gamma_d - 1) K_1^{1/2} = d^{\frac{1}{3} - \lambda_d} \sqrt{\ell K_1} \ .$$

Thus,

$$\lim_{d\to\infty} \mathbb{E}[|\mathcal{A}_{1,d}|] \leqslant \begin{cases} \sqrt{\ell K_1} < \infty & \text{if } \lim\limits_{d\to\infty} \lambda_d = 1/3 \ , \\ 0 & \text{if } \lim\limits_{d\to\infty} \lambda_d > 1/3 \ . \end{cases}$$

For $i = 2$, using (B.9) with $\lim\limits_{d\to\infty} K_{2,d} < \infty$, the limit of the first term of (B.11) becomes

$$\frac{\sigma_d^2 \sqrt{d}}{2!} \mathbb{E}[C_{2,1}^2]^{1/2} = \frac{\sigma_d^2 \sqrt{d}}{2} (\gamma_d - 1) \sqrt{K_{2,d}} = d^{\frac{1}{6}-\lambda_d} \frac{\ell^2}{2} \sqrt{K_{2,d}} \xrightarrow{d\uparrow\infty} 0 \ .$$

For the second term of (B.11), we find from (B.8) that

$$\frac{\sigma_d^2}{2!} d |\mathbb{E}[C_{2,1}]| \leqslant \frac{\sigma_d^2}{2} d (\gamma_d - 1)^2 \mathbb{E}[l'(X_1)^2] = d^{\frac{2}{3}-2\lambda_d} \frac{\ell^2}{2} \mathbb{E}[l'(X_1)^2] \ ,$$

with $\mathbb{E}[l'(X_1)^2] < \infty$ by the polynomial bound assumption. This means that

$$\lim_{d\to\infty} \mathbb{E}[|\mathcal{A}_{2,d}|] \leqslant \begin{cases} \frac{\ell^2}{2} \mathbb{E}[l'(X_1)^2] < \infty & \text{if } \lim\limits_{d\to\infty} \lambda_d = 1/3, \\ 0 & \text{if } \lim\limits_{d\to\infty} \lambda_d > 1/3. \end{cases}$$

For $i = 4$, since the second moment of $C_{4,1}$ is bounded, the first term of (B.11) satisfies

$$\frac{\sigma_d^4}{4!} \sqrt{d} \mathbb{E}[C_{4,1}^2]^{1/2} = d^{-\frac{1}{6}} \frac{\ell^4}{4!} \mathbb{E}[C_{4,1}^2]^{1/2} \xrightarrow{d\uparrow\infty} 0.$$

For the second term, we use $\mathbb{E}[C_{4,1}] = (\gamma_d - 1) K_{4,d}$ in (B.8) with $\lim\limits_{d\to\infty} K_{4,d} = L_4 < \infty$. This term is thus be bounded by

$$\frac{\sigma_d^4}{4!} d |\mathbb{E}[C_{4,1}]| \leqslant \frac{\sigma_d^4}{4!} d (\gamma_d - 1) |K_{4,d}| = d^{\frac{1}{3}-\lambda_d} \frac{\ell^4}{4!} |K_{4,d}|.$$

This implies that

$$\lim_{d\to\infty} \mathbb{E}[|\mathcal{A}_{4,d}|] \leqslant \begin{cases} \frac{\ell^4}{4!} |L_4| < \infty & \text{if } \lim\limits_{d\to\infty} \lambda_d = 1/3 \ , \\ 0 & \text{if } \lim\limits_{d\to\infty} \lambda_d > 1/3 \ . \end{cases}$$

Finally, using Jensen's inequality to bound $|\mathcal{A}_{6,d}|$ yields

$$\lim_{d\to\infty} \mathbb{E}[|\mathcal{A}_{6,d}|] \leqslant \lim_{d\to\infty} \frac{d\sigma_d^6}{6!} \mathbb{E}[|C_{6,1}|] = \lim_{d\to\infty} \frac{\ell^6}{6!} \mathbb{E}[|C_{6,1}|] \leqslant \lim_{d\to\infty} \frac{\ell^6}{6!} \mathbb{E}[C_{6,1}^2]^{1/2} < \infty \ .$$

Since $\lim\limits_{d\to\infty} \mathbb{E}[|R_d|] \leqslant \lim\limits_{d\to\infty} \sum\limits_{i=1}^{6} \mathbb{E}[|\mathcal{A}_{i,d}|] + \mathbb{E}[|U_d|] < \infty$, we conclude by Lemma B.6 that

$$\lim_{d\to\infty} \mathbb{E}[\alpha(X,Y)] > 0.$$

($\Leftarrow$) We need to show that if $\lim\limits_{d\to\infty} \lambda_d < 1/3$, then $\lim\limits_{d\to\infty} \mathbb{E}[\alpha(X,Y)] = 0$. From the second part of Lemma B.6, it is sufficient to verify that

$$\mathbb{E}[R_d] \xrightarrow{d\uparrow\infty} -\infty \qquad \text{and} \qquad \frac{\mathbb{E}[|R_d - \mathbb{E}[R_d]|]}{-\mathbb{E}[R_d]} \xrightarrow{d\uparrow\infty} 0 \ . \tag{B.12}$$

For the first limit, since $\mathbb{E}[C_{1,1}] = 0$, the dominating term in the development of $\mathbb{E}[R_d]$ is $\mathbb{E}[\mathcal{A}_{2,d}]$. Using (B.8), we have

$$\mathbb{E}[\mathcal{A}_{2,d}] = \frac{d\sigma_d^2}{2!}\mathbb{E}[C_{2,1}] = \frac{-(\gamma_d - 1)^2 d\sigma_d^2}{2}\mathbb{E}[l'(X_1)^2] = -d^{\frac{2}{3}-2\lambda_d}\frac{\ell^2}{2}\mathbb{E}[l'(X_1)^2] < 0 \ .$$

It follows that $\mathbb{E}[R_d] \to -\infty$ at a speed of $d^{\frac{2}{3}-2\lambda_d}$.

For the ratio in (B.12), we want to show that the numerator does not grow faster than the denominator. As shown in (B.4), we can bound $\mathbb{E}[|R_d - \mathbb{E}[R_d]|]$ by

$$\begin{aligned}
\mathbb{E}[|R_d - \mathbb{E}[R_d]|] &\leqslant \sum_{i=1}^{6} \frac{\sigma_d^i}{i!}\sqrt{d}\,\mathbb{V}(C_{i,1})^{1/2} + \mathbb{E}[|U_d - \mathbb{E}[U_d]|] \\
&= \sum_{i=1}^{6} \frac{\ell^i}{i!} d^{\frac{1}{2}-\frac{i}{6}}\mathbb{V}(C_{i,1})^{1/2} + \mathcal{O}(d^{-1/6}) \\
&= \ell d^{\frac{1}{3}}\mathbb{E}[C_{1,1}^2]^{1/2} + \frac{\ell^2}{2} d^{\frac{1}{6}}(\mathbb{E}[C_{2,1}^2] - \mathbb{E}[C_{2,1}]^2)^{1/2} + \mathcal{O}(1) \ ,
\end{aligned} \tag{B.13}$$

as the elements $i \geqslant 3$ of the sum are $\mathcal{O}(1)$ (because $\lim\limits_{d\to\infty} \mathbb{E}[C_{i,1}] < \infty$ and $\lim\limits_{d\to\infty} \mathbb{E}[C_{i,1}^2] < \infty$ for $i = 1, \ldots, 6$) and $\mathbb{E}[|U_d - \mathbb{E}[U_d]|]$ is $\mathcal{O}(d^{-1/6})$ as seen in the first part of this demonstration. Using (B.8) and (B.9), we simplify (B.13) and obtain

$$\begin{aligned}
&\mathbb{E}[|R_d - \mathbb{E}[R_d]|] \\
&\leqslant \ell d^{\frac{1}{3}}(\gamma_d - 1)K_1^{1/2} + \frac{\ell^2}{2} d^{\frac{1}{6}} \left\{(\gamma_d - 1)^2 K_{2,d} - (\gamma_d - 1)^4 \mathbb{E}[l'(X_1)^2]^2\right\}^{1/2} + \mathcal{O}(1) \\
&= d^{\frac{1}{3}-\lambda_d}\ell\sqrt{K_1} + d^{\frac{1}{6}-\lambda_d}\frac{\ell^2}{2}\left\{K_{2,d} - d^{-2\lambda_d}\mathbb{E}[l'(X_1)^2]^2\right\}^{1/2} + \mathcal{O}(1) \ .
\end{aligned}$$

This implies that $E[|R_d - \mathbb{E}[R_d]|]$ is $\mathcal{O}(d^{\frac{1}{3}-\lambda_d})$ while $\mathbb{E}[R_d]$ is $\mathcal{O}(d^{\frac{2}{3}-2\lambda_d})$. The ratio is thus $\mathcal{O}(d^{\lambda_d - \frac{1}{3}})$ and converges to 0 as $d \uparrow \infty$ since $\lim\limits_{d\to\infty} \lambda_d < 1/3$. By Lemma B.6, we conclude that $\lim_{d\to\infty} \mathbb{E}[\alpha(X,Y)] = 0$. □

References

Bédard, M. (2007). Weak convergence of Metropolis algorithms for non-i.i.d. target distributions. *Ann. Appl. Probab. 17*(4), 1222–1244.

Beskos, A. and A. Stuart (2009). MCMC methods for sampling function space. In *Invited Lectures, Sixth International Congress on Industrial and Applied Mathematics, ICIAM07, Editors Rolf Jeltsch and Gerhard Wanner*, pp. 337–364.

Boisvert-Beaudry, G. (2019). Efficacit des distributions instrumentales en quilibre dans un algorithme de type Metropolis-Hastings. Thesis, Université de Montréal.

Ethier, S. N. and T. G. Kurtz (1986). *Markov Processes: Characterization and Convergence.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. New York: John Wiley & Sons Inc.

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.

Sherlock, C. (2006). *Methodology for inference on the Markov modulated Poisson process and theory for optimal scaling of the random walk Metropolis.* Ph. D. thesis, Lancaster University.

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