Hierarchical models: Local proposal variances for RWM-within-Gibbs and MALA-within-Gibbs

Mylène Bédard[1,*]

Abstract

The performance of RWM- and MALA-within-Gibbs algorithms for sampling from hierarchical models is studied. For the RWM-within-Gibbs, asymptotically optimal tunings for Gaussian proposal distributions featuring a diagonal covariance matrix are developed using existing scaling analyses. This leads to locally optimal proposal variances that depend on the mixing components of the hierarchical model and that correspond to the classical asymptotically optimal acceptance rate of 0.234. Ignoring the local character of the optimal scaling is possible, leading to an optimal proposal variance that remains fixed for the duration of the algorithm; the corresponding asymptotically optimal acceptance rate is then shown to be lower than 0.234. Similar ideas are applied to MALA-within-Gibbs samplers, leading to efficient yet computationally affordable algorithms. Simplifications for location and scale hierarchies are presented, and findings are illustrated through numerical studies. The local and fixed approaches for the RWM- and MALA-within-Gibbs are compared to competitive samplers in the literature.

Keywords: efficiency, Gaussian proposal distribution, inhomogeneous proposal variances, Langevin diffusion, location/scale parameters, MALA

1. Introduction

The Random walk Metropolis algorithm (RWM) and the Metropolis-adjusted Langevin algorithm (MALA) are commonly used to produce samples from arbitrary distributions π that may be complex, high-dimensional, or both ([8]). The idea is to build a Markov chain $\{\mathbf{X}[j], j \in \mathbb{N}\}$ on a state space $\mathcal{X}$ by proposing candidates to be included in the process according to some acceptance probability. The resulting Markov chain admits the n-dimensional target distribution π as its unique invariant distribution. Hereafter, π shall also be used for denoting the target density on a state space $\mathcal{X}$ with respect to Lebesgue measure.

Suppose that the time-j state of the Markov chain is $\mathbf{X}[j] = \mathbf{x}$. In a (symmetrical) RWM algorithm for instance, the proposal distribution selected to generate a candidate $\mathbf{Y}[j+1] = \mathbf{y}$ for the next state of the chain is assumed to have a density $q_n(\mathbf{y}; \mathbf{x}) = q_n(|\mathbf{y} - \mathbf{x}|)$ with respect to Lebesgue measure. A pragmatic choice, on which we focus in this article, is to draw candidates from a $\mathcal{N}(\mathbf{x}, \sigma^2 I_n)$ for some $\sigma > 0$, where I_n is the n-dimensional identity matrix (the specific normal proposal distribution used with MALA shall be described in Section 5). In implementing RWM and MALA samplers, one can update all n components simultaneously (classical RWM/MALA), or divide them into subgroups to be updated consecutively (RWM- or MALA-within-Gibbs). The latter are commonly preferred for sampling hierarchical models, as full conditional densities are usually available.

The variance of the normal proposal distribution (σ^2) has a significant impact on the speed at which the process travels across its state space (hereafter referred to as "efficiency"), with extremal variances leading to slow-mixing samplers. Simply put, large variances induce lazy

*Corresponding author
Email address: bedard@dms.umontreal.ca (Mylène Bédard)
[1]Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, H3C 3J7, Canada.

processes (large candidate jumps that are refused), while small variances yield hyperactive processes (tiny candidate steps that are accepted). To optimize exploration of the state space, we aim for candidate steps that strike a balance, so that we have sizable steps that are still accepted a reasonable proportion of the time. Seeking for this intermediate proposal variance is called the optimal scaling problem.

There exist, in Markov chain Monte Carlo theory, different notions of efficiency. In this paper, the term efficiency is used as a measure of how rapidly the Markov chain explores its state space once stationarity has been reached. For finite-dimensional chains, this can be measured by the expected squared jumping distance (ESJD) to be introduced in (13). In an infinite-dimensional setting, the theoretical (or asymptotical) efficiency is measured through the speed function of the limiting Langevin diffusion, to be discussed in Sections 3 and 4. In the high-dimensional limit ($n \to \infty$), the ESJD is equivalent to the limiting speed measure.

This paper studies the optimal scaling theory for RWM-within-Gibbs with some heuristics for MALA-within-Gibbs, and then looks at the performance of both in practice. In particular, the theory exposed leads to the determination of proposal variances and acceptance rates producing optimally mixing RWM-within-Gibbs chains. The theoretical results are derived for high-dimensional hierarchical target densities with a large number of conditionally independent and identically distributed (i.i.d.) components. The principal difference with traditional optimal scaling results lies in the local character of the optimal proposal variances obtained, meaning that they vary from one iteration to the next. The concept of local proposal variances has been discussed in [6] and [2]; in the latter, scaling analyses of the RWM algorithm for hierarchical target densities are performed. Although theoretically appealing, local proposal variances had to be obtained numerically in that context, which turned out to be rather impractical. With the RWM-within-Gibbs sampler (and even the MALA-within-Gibbs), these variances may now be found analytically in a large number of cases, leading to a personalized version of the proposal variance in a given iteration. The theoretical results derived thus stand on the work in [2], and as such are expressed as a corollary of its main theorem.

The derivation of local proposal variances requires that certain expectations be obtained analytically from the hierarchical model considered. The new approach is thus predicated on the tractability of the distribution of the conditionally i.i.d. components, given the mixing parameters and (in practice) the observations. It is thus well suited to some hierarchical models; alternatively, we propose a fixed optimal proposal variance, which is shown to be less efficient than the local ones. In an attempt to quantify the benefit, in terms of efficiency, of using local proposal variances rather than a fixed one in the RWM- and MALA-within-Gibbs, we present numerical illustrations. To add some perspective, we compare these samplers to single-block RWM and MALA algorithms, along with some of their variants that include correlation among candidates. We also include the Adaptive Metropolis (AM) sampler of [7], which tunes the proposal covariance matrix on the fly.

We shall realize that in tractable cases (and when there is not a strong correlation between mixing parameters and the remaining components), local versions of RWM- and MALA-within-Gibbs can outperform fancy variants included in the MCMC toolbox. Local MALA-within-Gibbs is the approach that provides the most convincing results, leading to net efficiency gains in a wide range of situations, compared to a large set of competitors. These gains are however largely influenced by the degree of variability present in the hierarchical model (a large variability sustaining the pursuit of local proposal variances). Even in cases where local samplers do not allow for large gains in terms of theoretical efficiency, the risk associated with these local variances is limited to the extra computational effort required for their implementation, which is usually insignificant compared to a fixed variance.

The next section sets up the framework, while Section 3 reviews optimal scaling notions for

high-dimensional i.i.d. and hierarchical targets. Understanding these notions turns out to be useful in Section 4, where we derive optimal tunings for the RWM-within-Gibbs for sampling from hierarchical models; extensions to MALA-within-Gibbs are then discussed in Section 5. Section 6 focuses on single-level hierarchical models where the mixing parameter acts on the location or scale of the conditionally i.i.d. components; a simulation study illustrates the theoretical results. An extension to inhomogeneous proposal variances is introduced in Section 7, and we conclude by presenting a numerical study on a hierarchical target model that falls slightly outside the assumptions of the theory (Section 8).

2. Framework

Consider the following $(n+p)$-dimensional target density π with respect to Lebesgue measure

$$\pi(\mathbf{x}^{(n+p)}) \quad = \quad f_1(x_1, \ldots, x_p) \prod_{i=p+1}^{p+n} f(x_i|x_1, \ldots, x_p) \; ; \tag{1}$$

this is a multi-level hierarchical model with p mixing components $\mathbf{X}_{1:p} = (X_1, \ldots, X_p)$ and n conditionally i.i.d. components $\mathbf{X}_{(p+1):(p+n)}$ given $\mathbf{X}_{1:p}$. We impose some regularity conditions ensuring that the density f is smooth on $\mathbb{R}$. Let $\mathcal{X}_1 = \{\mathbf{x}_{1:p} : f_1(\mathbf{x}_{1:p}) > 0\}$; for fixed $\mathbf{x}_{1:p} \in \mathcal{X}_1$, $f(x|\mathbf{x}_{1:p})$ is a positive C^2 density on $\mathbb{R}$ (C^2 denotes the space of real-valued functions with continuous second derivatives). Furthermore, for all fixed $\mathbf{x}_{1:p} \in \mathcal{X}_1$, $\frac{\partial}{\partial x} \log f(x|\mathbf{x}_{1:p})$ is Lipschitz continuous with constant $k(\mathbf{x}_{1:p})$ such that $\mathbb{E}[k^4(\mathbf{X}_{1:p})] < \infty$, and

$$\mathbb{E}_X\left[\left(\tfrac{\partial}{\partial X} \log f(X|\mathbf{x}_{1:p})\right)^4\right] < \infty \; \forall \mathbf{x}_{1:p} \in \mathcal{X}_1 \quad \text{with} \quad \mathbb{E}\left[\left(\tfrac{\partial}{\partial X} \log f(X|\mathbf{X}_{1:p})\right)^4\right] < \infty \; .$$

Hereafter, the notation $\mathbb{E}_X[\cdot]$ means that the expectation is computed with respect to $X|\mathbf{x}_{1:p}$ with density $f(x|\mathbf{x}_{1:p})$, while $\mathbb{E}[\cdot]$ is used to denote an expectation with respect to all the random variables in the expression (so $X|\mathbf{x}_{1:p}$ as before, but also $\mathbf{X}_{1:p}$ with density f_1).

To sample from such target densities, consider the following RWM-within-Gibbs algorithm with Gaussian proposal distributions. If the $(p+n)$ target components in (1) are divided into two blocks of dimensions p and n respectively, *i.e.* $(\mathbf{X}_1, \mathbf{X}_2) = (\mathbf{X}_{1:p}, \mathbf{X}_{(p+1):(p+n)})$, then one iteration is implemented as follows.

Algorithm 1 (RWM-within-Gibbs algorithm).
1) Given the time-j state $\mathbf{X}[j] = (\mathbf{X}_1[j], \mathbf{X}_2[j]) = (\mathbf{x}_1, \mathbf{x}_2)$ of the Markov chain, generate a p-dimensional candidate $\mathbf{Y}_1[j+1]$ for the first block according to

$$\mathbf{Y}_1[j+1] = \mathbf{x}_1 + \sigma_1 \; \mathbf{Z}_1[j+1] \; ,$$

where $\sigma_1 > 0$ and $\mathbf{Z}_1[j+1] \sim \mathcal{N}(0, I_p)$.

2) Accept the candidate $\mathbf{Y}_1[j+1] = \mathbf{y}_1$ with probability

$$\alpha(\mathbf{x}_1; \mathbf{y}_1|\mathbf{x}_2) = 1 \wedge \frac{\pi(\mathbf{y}_1|\mathbf{x}_2)}{\pi(\mathbf{x}_1|\mathbf{x}_2)} \; ,$$

where $\pi(\mathbf{x}_1|\mathbf{x}_2) \propto \pi(\mathbf{x}^{(n+p)})$ is the conditional density of $\mathbf{x}_1$ given that $\mathbf{X}_2[j] = \mathbf{x}_2$. That is, $\mathbf{X}_1[j+1] = \mathbf{y}_1$ with probability $\alpha(\mathbf{x}_1; \mathbf{y}_1|\mathbf{x}_2)$, and $\mathbf{X}_1[j+1] = \mathbf{x}_1$ otherwise.

3) Given the updated state of the first block, $\mathbf{X}_1[j+1] = \mathbf{x}_1^*$, generate an n-dimensional candidate $\mathbf{Y}_2[j+1]$ for the second block according to

$$\mathbf{Y}_2[j+1] = \mathbf{x}_2 + \frac{\ell}{\sqrt{n}} \; \mathbf{Z}_2[j+1] \; ,$$

where $\ell = \ell(\mathbf{x}_1^*) > 0$ *and* $\mathbf{Z}_2[j+1] \sim \mathcal{N}(0, I_n)$.

4) Accept the candidate $\mathbf{Y}_2[j+1] = \mathbf{y}_2$ *with probability*

$$\alpha(\mathbf{x}_2; \mathbf{y}_2 | \mathbf{x}_1^*) = 1 \wedge \frac{\pi(\mathbf{y}_2 | \mathbf{x}_1^*)}{\pi(\mathbf{x}_2 | \mathbf{x}_1^*)} \ ,$$

where $\pi(\mathbf{x}_2 | \mathbf{x}_1^*) \propto \pi((\mathbf{x}_1^*, \mathbf{x}_2))$ *is the conditional density of* $\mathbf{x}_2$ *given that* $\mathbf{X}_1[j+1] = \mathbf{x}_1^*$. *If the candidate is accepted, set* $\mathbf{X}_2[j+1] = \mathbf{y}_2$*; otherwise,* $\mathbf{X}_2[j+1] = \mathbf{x}_2$.

This sampler may easily be generalized to a larger number of blocks; the particular case in which there is only one block of components leads to the single-block RWM algorithm. Scaling analyses of the latter have already been performed for target densities such as (1) (see [2]). In the next section, we describe some optimal scaling results that have been introduced in the statistical literature, and that turn out to be useful in the tuning of the RWM-within-Gibbs sampler.

3. Available literature about optimal scaling

The optimal scaling issue of the RWM algorithm with a Gaussian proposal has been addressed by many researchers over the last few decades. The seminal work of [15] presents a solution for tuning these algorithms when the target density is formed of n i.i.d. components, where n is large.

Generalizing the theoretical result of [15] is an intricate task; further research on this subject has addressed the tuning of various algorithms, but has mainly been restricted to the case of high-dimensional target distributions formed of independent components (see [13], [12], [11], [1], [5], [3], and the references therein). A few of these papers have extended their tuning results for multivariate normal targets with correlation. Lately, scaling analyses of the RWM sampler for non-product target densities have also been performed in [4], [16], [10], and [2].

3.1. Optimal tuning of isotropic RWM: product target densities

Consider a RWM algorithm with an isotropic proposal distribution $\mathcal{N}(\mathbf{X}^{(n)}[j], \sigma^2(n) I_n)$, applied to a product target density $\pi(\mathbf{x}^{(n)}) = \prod_{i=1}^n f(x_i)$ with respect to Lebesgue measure. The one-dimensional density f satisfies the assumptions specified in Section 2 (without mixing components). It is well understood by now that the regularity conditions stated in [15] are stronger than necessary; those in Section 2 are slightly different, but sufficient for the following theorem to hold according to [2].

Asymptotically optimal scaling results are obtained by studying the limiting path (as $n \to \infty$) of a given component ($X_1^{(n)}$ say). The one-dimensional process is studied conditionally on $\mathcal{F}^{\mathbf{X}^{(n)}}[j]$, the filtration of the n-dimensional process up to time j (the current state). In order to obtain a non-trivial limiting process as $n \to \infty$, space and time rescaling factors must be applied to the RWM algorithm. The proposal variance is set to be a decreasing function of the dimension by letting $\sigma^2(n) = \ell^2/n$. A continuous-time sped-up version of the initial Markov chain is then introduced as $\{\mathbf{W}^{(n)}(t); t \geq 0\} = \{\mathbf{X}^{(n)}[\lfloor nt \rfloor]; t \geq 0\}$, where $\lfloor \cdot \rfloor$ is the floor function. Note that $[\cdot]$ and $(\cdot)$ are respectively used to index time in discrete- and continuous-time versions of the Markov chain.

Hereafter, let $\Rightarrow$ denote weak convergence in the Skorokhod topology and $B(t)$ a Brownian motion at time t; the cumulative distribution function of a standard normal random variable is denoted by $\Phi(\cdot)$. The theoretical result proven in [15] for tuning RWM algorithms with an isotropic proposal applied to product target densities is stated below.

Theorem 2. *Suppose that* $\mathbf{W}^{(n)}(0)$ *is distributed according to* $\pi(\mathbf{x}^{(n)})$. *For* $i = 1, 2, \ldots$, *we have* $\{W_i^{(n)}(t); t \geq 0\} \Rightarrow \{W_i(t); t \geq 0\}$, *where* $W_i(0)$ *is distributed according to the density* f, *and* $\{W_i(t); t > 0\}$ *satisfies the stochastic differential equation (SDE)*

$$dW_i(t) = \upsilon^{1/2}(\ell)\ dB(t) + \frac{1}{2}\ \upsilon(\ell)\ \tfrac{d}{dW_i(t)} \log f\left(W_i(t)\right) dt,$$

with

$$\upsilon(\ell) = 2\ell^2 \Phi\left(-\frac{\ell}{2}\sqrt{\mathbb{E}\left[\left(\tfrac{d}{dX}\log f(X)\right)^2\right]}\right). \tag{2}$$

Furthermore, the univariate processes $\{W_i^{(n)}(t); t \geq 0\}$, $i = 1, 2, \ldots, n$, *are asymptotically mutually independent as* $n \to \infty$.

The components of the isotropic RWM algorithm thus asymptotically behave according to mutually independent Langevin diffusion processes, which depend on ℓ through the speed function $\upsilon(\ell)$. The Langevin diffusion process that travels the most rapidly across its state space is the process for which $\upsilon(\ell)$ is maximized with respect to ℓ, giving rise to the asymptotically optimal scaling value $\hat{\ell}$ for the algorithm. The optimal proposal variance satisfies $\hat{\sigma}^2(n) = \hat{\ell}^2/n \approx 5.66/\{n\mathbb{E}\left[\{\log f(X))'\}^2\right]\}$ and corresponds to an optimal expected acceptance rate of 23.4%, where the acceptance rate is defined as the proportion of candidates that are accepted by the algorithm.

3.2. Optimal tuning of RWM: hierarchical target densities

In [2], asymptotically optimal scaling results similar to those of [15] are derived, but for target densities as in (1) with $p = 1$. To prove the theoretical results, some further regularity assumptions are imposed on $f(x|x_1)$, for fixed $x \in \mathbb{R}$. The only assumption on f_1 is that it be a continuous density, with $\mathcal{X}_1$ forming an open interval on $\mathbb{R}$; we refer the reader to [2] for more details.

Consider a RWM algorithm with proposal distribution $\mathcal{N}(\mathbf{x}, \ell^2 I_{n+1}/(n+1))$; that is, no prior knowledge about the target correlation structure is assumed. Although not a concern in the i.i.d. setting, we now emphasize the fact that the proposal variance $\sigma^2(n) = \ell^2/(n+1)$ may be a function of the current state $\mathbf{x}$ through $\ell = \ell(x_1)$. Such a variation results in a valid non-homogeneous RWM sampler, provided that the ratio of proposal densities be included in the acceptance probability : $\alpha(\mathbf{x}; \mathbf{y}) = 1 \wedge \pi(\mathbf{y}) q_{n+1}(\mathbf{x}; \mathbf{y}) / \{\pi(\mathbf{x}) q_{n+1}(\mathbf{y}; \mathbf{x})\}$, where $q_{n+1}(\mathbf{y}; \mathbf{x})$ is the density of a $\mathcal{N}(\mathbf{x}, \ell^2(\mathbf{x}) I_{n+1}/(n+1))$.

By studying an appropriately rescaled version of the initial Markov process, [2] obtains a weak convergence result pointing towards the use of proposal variances that are a function of x_1, the current state of the mixing component. Specifically, a local proposal variance which is set to $\hat{\sigma}^2(x_1, n) = \hat{\ell}^2(x_1)/(n+1)$ maximizes the speed function

$$\upsilon(\ell, x_1) = 2\ell^2 \mathbb{E}\left[\Phi\left(-\frac{\ell}{2}\gamma^{1/2}(x_1, Z_1)\right)\right], \tag{3}$$

with respect to ℓ. Here, $Z_1 \sim \mathcal{N}(0, 1)$ and

$$\gamma(x_1, z_1) \quad = \quad z_1^2\ \mathbb{E}_X\left[\left(\tfrac{\partial}{\partial x_1}\log f(X|x_1)\right)^2\right] + \mathbb{E}_X\left[\left(\tfrac{\partial}{\partial X}\log f(X|x_1)\right)^2\right]. \tag{4}$$

The speed function (3) is intuitive: the second term in (4) is a measure of roughness of $f(x|x_1)$ under a variation of x, and is similar to the quantity $\mathbb{E}[\{(\log f(X))'\}^2]$ in the i.i.d.

case. With hierarchical targets, we find an extra term that might be viewed as a measure of roughness of $f(x|x_1)$ under a variation of x_1. This term is weighted by z_1^2, the squared candidate standardized increment for the first component. The speed function in (3) then averages over the possible standardized increments z_1.

Although the concept of local tunings is theoretically appealing, the weak convergence results in [2] were carried with fixed ℓ; it is thus unclear whether local proposal variances optimizing (3) really are conditionally optimal given x_1. The proof of such a result would require including the ratio of proposal densities in the acceptance probability of the sampler, which is not a trivial extension. These values are also difficult to obtain in practice: $\hat{\ell}(x_1)$ must be obtained numerically and updated every time X_1 jumps to a new state. Since the process is assumed to start in stationarity (as in Theorem 2), we overcome this difficulty by defining an asymptotically optimal scaling value $\hat{\ell}$ that is fixed for the duration of the algorithm. This value maximizes the expectation of the speed function with respect to the marginal distribution of X_1,

$$\mathbb{E}\left[\upsilon\left(\ell, X_1\right)\right] \quad = \quad 2\ell^2 \int_{\mathcal{X}_1} \int_{\mathbb{R}} \Phi\left(-\frac{\ell}{2}\gamma^{1/2}(x_1, z_1)\right) \phi(z_1) f_1(x_1)\, dz_1\, dx_1,$$

where $\phi(\cdot)$ is the probability density function of a standard normal random variable. In the current context the fixed optimal scaling value, which results from a single (numerical) maximization performed before running the algorithm, is thus preferred to local tunings. In subsequent examples, we shall automatically consider fixed optimal scaling values when dealing with single-block RWM algorithms.

4. Tuning the RWM-within-Gibbs

Using the asymptotic results described in Section 3, we now optimize the efficiency of RWM-within-Gibbs samplers (Algorithm 1) applied to hierarchical target models. For simplicity, we update each of the p mixing components in turn using a $\mathcal{N}(X_i[j], \sigma_i^2)$ for $i = 1, \ldots, p$ since it is then unnecessary to estimate the correlation structure. Parameters in a hierarchical model may however be highly correlated, in which case it might be better to update them as a single block, using an estimate of the covariance matrix; appropriate transformations may also be applied to parameters that are initially restricted to subsets of $\mathbb{R}$. We then take advantage of conditional independence among the components $\mathbf{X}_{(p+1):(p+n)}$ (given $\mathbf{X}_{1:p}$) by grouping them into a block. Given the updated states of the mixing components at time $j+1$, $\mathbf{X}_{1:p}[j+1] = \mathbf{x}_{1:p}^*$, a candidate $\mathbf{Y}_{(p+1):(p+n)}[j+1]$ is generated using a $\mathcal{N}(\mathbf{X}_{(p+1):(p+n)}[j], \sigma^2(\mathbf{x}_{1:p}^*, n) I_n)$, with $\sigma^2(\mathbf{x}_{1:p}^*, n) = \ell^2(\mathbf{x}_{1:p}^*)/n$. Under this setting, the last n components are updated simultaneously, but independently from each other, using a proposal variance that may vary from one iteration to another.

For a one-dimensional normal target, it is widely known that an optimally efficient version of the RWM algorithm should be tuned to accept approximately 45% of the proposed candidates (see [12], [16]). We shall then apply this result to tune the first p proposal variances, $\sigma_1^2, \ldots, \sigma_p^2$. In the subsequent examples, tunings of the different blocks are performed simultaneoulsy and independently, as the scaling of a particular block does not affect the tuning of the other blocks. When this is not the case, it might again be more efficient to update all p mixing components in a single block. To obtain asymptotically optimal scaling results for updating the last block of the RWM-within-Gibbs sampler, we assume that p is fixed; the number of components in the last block thus grows as $n \to \infty$. Based on previous optimal scaling theory, relying on the process $\{\mathbf{W}^{(p+n)}(t); t \geq 0\} = \{\mathbf{X}^{(p+n)}[\lfloor nt \rfloor]; t \geq 0\}$ (which is the continuous-time, sped-up version of the initial Markov chain described in Section 3.1) along with a proposal variance that decreases linearly in n, shall lead to a non-trivial limiting process as n grows.

Even though the conditional density of $\mathbf{X}_{(p+1):(p+n)}$ given $\mathbf{X}_{1:p}$ is expressed as a product, $\prod_{i=p+1}^{p+n} f(x_i|\mathbf{x}_{1:p})$, one cannot directly rely on the results in [15]. Indeed, we are dealing with a density whose shape varies with the same rhythm as $\mathbf{X}_{1:p}$ is updated. To tune this sampler, we have a choice between using local or constant proposal variances. The latter is the usual variance that is fixed for the duration of the algorithm, expressed as ℓ^2/n, while the former consists in a variance that depends on $\mathbf{X}_{1:p}[j+1] = \mathbf{x}^*_{1:p}$, *i.e.* that is updated at every iteration. In the current context, relying on local proposal variances does not require an adjustment of the acceptance probability since the p mixing parameters are updated separately.

4.1. Local proposal variances

In a given iteration, the components $\mathbf{X}_{1:p}$ are updated prior to the block $\mathbf{X}_{(p+1):(p+n)}$. Given $\mathbf{W}^{(p+n)}(t)$, the time-t value of the rescaled process, denote the next instant by $t+dt$ and let the updated value of $\mathbf{X}_{1:p}$ at $t+dt$ be expressed as $\mathbf{W}^{(p+n)}_{1:p}(t+dt) = \mathbf{x}^*_{1:p}$. Because of this specificity of the RWM-within-Gibbs sampler, it is possible to study the limiting behaviour of the components in the last block by conditioning on the filtration $(\mathcal{F}^{\mathbf{W}^{(p+n)}_{1:p}}(t+dt), \mathcal{F}^{\mathbf{W}^{(p+n)}_{(p+1):(p+n)}}(t))$. We then suppose that the proposal variance of the last block satisfies $\sigma^2(n, \mathbf{x}^*_{1:p}) = \ell^2(\mathbf{x}^*_{1:p})/n < \infty$ for all $\mathbf{x}^*_{1:p} \in \mathcal{X}_1$, with $\mathbb{E}[\ell^{12}(\mathbf{X}^*_{1:p})] < \infty$. Taking advantage of the conditional independence among the components $\mathbf{X}_{(p+1):(p+n)}$ given $\mathbf{X}_{1:p}$ and applying arguments similar to those in [2] lead to the following result.

Corollary 3. *Suppose that $\mathbf{W}^{(n+p)}(0)$ is distributed according to π in (1) and let $\ell \equiv \ell(\mathbf{x}^*_{1:p}) > 0$. For $i = p+1, \ldots, p+n$, we have $\{W_i^{(p+n)}(t); t \geq 0\} \Rightarrow \{W_i(t); t \geq 0\}$, where $\mathbf{W}_{1:p}(0)$ and $W_i(0)$ are distributed according to the densities f_1 and f respectively. Conditionally on the filtration $(\mathcal{F}^{\mathbf{W}_{1:p}}(t+dt), \mathcal{F}^{W_i}(t))$, the evolution of the process $\{W_i(t); t > 0\}$ may be expressed as*

$$\begin{aligned} dW_i(t) &= \upsilon^{1/2}(\ell, \mathbf{W}_{1:p}(t+dt))\; dB(t) \\ &\quad + \frac{1}{2}\; \upsilon(\ell, \mathbf{W}_{1:p}(t+dt))\; \tfrac{\partial}{\partial W_i(t)} \log f\left(W_i(t)|\mathbf{W}_{1:p}(t+dt)\right) dt, \end{aligned} \tag{5}$$

with

$$\upsilon(\ell, \mathbf{x}^*_{1:p}) = 2\ell^2 \Phi\left(-\frac{\ell}{2}\sqrt{\mathcal{I}(\mathbf{x}^*_{1:p})}\right), \tag{6}$$

and $\mathcal{I}(\mathbf{x}^*_{1:p}) = \mathbb{E}_X\left[\left(\frac{\partial}{\partial X}\log f(X|\mathbf{x}^*_{1:p})\right)^2\right]$.

This means that conditionally on $\boldsymbol{W}_{1:p}(t+dt) = \mathbf{x}^*_{1:p}$, the evolution of $\{W_i(t); t > 0\}$ over an infinitesimal interval dt satisfies

$$\{W_i(t+dt) - W_i(t)\}|W_i(t), \mathbf{x}^*_{1:p} \sim \mathcal{N}\left(\tfrac{1}{2}\; \upsilon(\ell, \mathbf{x}^*_{1:p})\; \tfrac{\partial}{\partial W_i(t)} \log f\left(W_i(t)|\mathbf{x}^*_{1:p}\right) dt, \upsilon(\ell, \mathbf{x}^*_{1:p})\; dt\right)\; .$$

Proof. The result can be seen as a special case of Theorem 2 in [2], which describes a weak convergence result for the single-block RWM applied to hierarchical target models (Section 3.2). Indeed, since $\mathbf{X}_{1:p}$ and $\mathbf{X}_{(p+1):(p+n)}$ are updated successively, the last block of the RWM-within-Gibbs deals with a product target density that depends on the mixing components $\mathbf{X}_{1:p}$ (instead of a hierarchical target, which considerably simplifies the proof). Since these mixing components have already been updated, the process is unaffected by the roughness of the density $f(x_i|\mathbf{x}_{1:p})$ under a variation of $\mathbf{x}_{1:p}$. The first term in (4) is thus null, which implies that speed functions in (6) and (3) are equivalent.

We should however point out that the proof of Theorem 2 in [2] applies when ℓ is fixed, as local proposal variances are not practically appealing in the RWM framework. For the current RWM-within-Gibbs sampler, a weak convergence result with local proposal variances $\sigma^2(x_1, n) = \ell^2(x_1)/n$ nonetheless follows from the arguments in the proof of Theorem 2 of [2]. As mentioned previously, since $\mathbf{X}_{1:p}$ are updated separately, we may rely on local proposal variances without having to include the ratio of proposal densities in the acceptance probability. The arguments in the proof, which is based on the $\mathcal{L}^1$-convergence of generators, thus remain the same as before. The only adjustment when replacing ℓ by $\ell(\mathbf{x}^*_{1:p})$ lies in the control of the error terms, which is straight-forwardly achieved by making use of the regularity assumptions in Section 2 along with the Cauchy-Schwarz inequality. To this end, slightly stronger regularity conditions are imposed compared to [2]; in particular, we require $\mathbb{E}[k^4(\mathbf{X}_{1:p})] < \infty$ (instead of $\mathbb{E}[k^2(\mathbf{X}_{1:p})] < \infty$, where $k(\mathbf{x}_{1:p})$ is the Lispchitz constant defined in Section 2) and also that proposal variances be finite for all $\mathbf{x}^*_{1:p} \in \mathcal{X}_1$, with $\mathbb{E}[\ell^{12}(\mathbf{X}_{1:p})] < \infty$. □

The process $\{W_i(t); t \geq 0\}$ is Markovian with respect to the filtration $(\mathcal{F}^{\mathbf{W}_{1:p}}(t+dt), \mathcal{F}^{W_i}(t))$. The equation in (5) provides a picture of the asymptotic behaviour of that process at time t, once the first p components have been updated. By optimizing the speed measure of the limiting diffusion with respect to ℓ, we find the best possible proposal variance to use at time t for generating a candidate at time $t+dt$, given that $\mathbf{W}_{1:p}(t+dt) = \mathbf{x}^*_{1:p}$. Optimizing $\upsilon(\ell, \mathbf{x}^*_{1:p})$ with respect to ℓ leads to

$$\hat{\ell}(\mathbf{x}^*_{1:p}) = \frac{2.38}{\left\{\mathcal{I}(\mathbf{x}^*_{1:p})\right\}^{1/2}} \; ; \tag{7}$$

the function $\hat{\sigma}^2(\mathbf{x}^*_{1:p}, n) = \hat{\ell}^2(\mathbf{x}^*_{1:p})/n$ thus represents a locally optimal proposal variance.

We now introduce the expected acceptance rate of the n-dimensional stationary RWM step. To be concise, let $\mathbf{x}_{(p+1):(p+n)} = \mathbf{x}$ and $\mathbf{y}_{(p+1):(p+n)} = \mathbf{y}$; then,

$$\begin{aligned} a_n(\ell(\mathbf{x}^*_{1:p})) \;\; = \;\; & \iiint \alpha(\mathbf{x}; \mathbf{y} | \mathbf{x}^*_{1:p}) \left(\frac{\ell(\mathbf{x}^*_{1:p})}{\sqrt{n}}\right)^{-n} \phi_n\left(\frac{\mathbf{y}-\mathbf{x}}{\ell(\mathbf{x}^*_{1:p})/\sqrt{n}}\right) \\ & \qquad \prod_{i=p+1}^{p+n} f(x_i|\mathbf{x}^*_{1:p}) f_1(\mathbf{x}^*_{1:p}) \; d\mathbf{y} \; d\mathbf{x} \; d\mathbf{x}^*_{1:p} \; , \end{aligned}$$

where $\phi_n(\cdot)$ stands for the probability density function of an n-dimensional standard normal random variable. By reproducing the optimal scaling proofs in [2], it is easy to show that

$$\lim_{n\to\infty} a_n(\ell(\mathbf{x}^*_{1:p})) = a(\ell(\mathbf{x}^*_{1:p})) \equiv 2 \int \Phi\left(-\frac{\ell(\mathbf{x}^*_{1:p})}{2}\sqrt{\mathcal{I}(\mathbf{x}^*_{1:p})}\right) f_1(\mathbf{x}^*_{1:p}) \; d\mathbf{x}^*_{1:p} \; . \tag{8}$$

To find the asymptotically optimal acceptance rate corresponding to the locally optimal proposal variances, it suffices to evaluate (8) at $\hat{\ell}(\mathbf{x}^*_{1:p})$, which yields

$$a(\hat{\ell}(\mathbf{x}^*_{1:p})) \;\; = \;\; 2\Phi\left(-\frac{2.38}{2}\right) \int f_1(\mathbf{x}^*_{1:p}) \; d\mathbf{x}^*_{1:p} \approx 0.234 \; .$$

It is interesting to note that the asymptotically optimal acceptance rate of the RWM block update is still equal to 0.234. This optimally mixing subchain cannot, however, be obtained by monitoring the acceptance rate. Indeed, building an optimal Markov chain based on local proposal variances implies modifying the proposal variance at every iteration.

In contrast to the local proposal variances derived for single-block RWM algorithms in Section 3.2, an explicit expression for $\mathcal{I}(\mathbf{x}^*_{1:p})$ is often available in practice for RWM-within-Gibbs samplers. As shall be witnessed below, the local tuning approach is particularly useful when dealing with various models involving scale parameters. For hierarchical models related through parameters other than location or scale, the benefit of relying on this type of tuning requires a case by case analysis.

4.2. Fixed proposal variance

It is also possible to obtain an optimal, fixed proposal variance; the latter is however not the best proposal variance given $\mathbf{x}^*_{1:p}$, the latest update of the first p components. By assuming that the Markov chain starts in stationarity (as was done in the previous sections), the optimal fixed proposal variance, $\hat{\sigma}^2(n) = \hat{\ell}^2/n$, optimizes

$$\mathbb{E}\left[v(\ell, \mathbf{X}_{1:p})\right] = 2\ell^2 \int \Phi\left(-\frac{\ell}{2}\sqrt{\mathcal{I}(\mathbf{x}_{1:p})}\right) f_1(\mathbf{x}_{1:p})\, d\mathbf{x}_{1:p} \tag{9}$$

with respect to ℓ. A corresponding asymptotically optimal acceptance rate $a(\hat{\ell})$ may then be computed (both values are generally obtained numerically). Apart from specific frameworks, a fixed approach will generally not match the results arising from a local approach. We thus expect an acceptance rate no greater than 0.234, which is the acceptance rate obtained with the best possible combination of (local) proposal variances. In fact, an optimal fixed proposal variance only gives rise to an acceptance rate of 0.234 when $\hat{\ell}^2(\mathbf{x}^*_{1:p})$ is independent of $\mathbf{x}^*_{1:p}$, in which case the locally optimal proposal variance is constant.

Proposition 4. *Consider a RWM-within-Gibbs sampler and a target density as in Section 2. The asymptotic acceptance rate $a(\hat{\ell})$ obtained by applying the optimal fixed proposal variance $\hat{\sigma}^2(n)$ is no greater than 0.234, the asymptotically optimal acceptance rate arising from the locally optimal proposal variances in (7).*

Proof. The proof is similar to that of Theorem 4 in [16], except when we introduce g at the end, which corrects a sign typographic error in their proof. Since the optimal fixed proposal variance optimizes $\mathbb{E}[v(\ell, X_{1:p})]$ with respect to ℓ, we have

$$2\mathbb{E}\left[\Phi\left(-\frac{\hat{\ell}}{2}\sqrt{\mathcal{I}(\mathbf{X}_{1:p})}\right)\right] = \mathbb{E}\left[\frac{\hat{\ell}}{2}\sqrt{\mathcal{I}(\mathbf{X}_{1:p})}\phi\left(-\frac{\hat{\ell}}{2}\sqrt{\mathcal{I}(\mathbf{X}_{1:p})}\right)\right] .$$

Letting $t(\mathbf{X}_{1:p}) = \Phi(-\frac{\hat{\ell}}{2}\sqrt{\mathcal{I}(\mathbf{X}_{1:p})})$ with $t(\mathbf{X}_{1:p}) \in [0, 0.5]$ (since $\hat{\ell} \geq 0$ and $\sqrt{\mathcal{I}(\mathbf{X}_{1:p})} \geq 0$), this becomes

$$2\mathbb{E}\left[t(\mathbf{X}_{1:p})\right] = \mathbb{E}\left[-\Phi^{-1}\left(t(\mathbf{X}_{1:p})\right)\phi\left(\Phi^{-1}\left(t(\mathbf{X}_{1:p})\right)\right)\right] .$$

As argued in [16], we may apply Jensen's inequality to obtain

$$\mathbb{E}\left[-\Phi^{-1}\left(t(\mathbf{X}_{1:p})\right)\phi\left(\Phi^{-1}\left(t(\mathbf{X}_{1:p})\right)\right)\right] \leq -\Phi^{-1}\left(\mathbb{E}[t(\mathbf{X}_{1:p})]\right)\phi\left(\Phi^{-1}\left(\mathbb{E}[t(\mathbf{X}_{1:p})]\right)\right) .$$

Combining the previous two equations and expressing them in terms of $g = -\Phi^{-1}(\mathbb{E}[t(\mathbf{X}_{1:p})])$ yields $2\Phi(-g) \leq g\phi(-g)$. The single solution in the case of equality is $\hat{g} \approx 1.19$, and the inequality is strict if and only if $g > \hat{g}$; hence, the acceptance rate satisfies $2\Phi(g) \leq 2\Phi(\hat{g})$. □

In light of the results expounded in Section 4, one should rely on locally optimal proposal variances when available; these are efficient and computationally inexpensive to implement, as shall be seen in the coming sections. Alternatively, one may use the (optimal) fixed proposal variance or its corresponding acceptance rate for tuning the last block. The optimal acceptance rate is obtained as a function of the fixed proposal variance; this shall be illustrated in Sections 6.3 and 8.

5. MALA-within-Gibbs

The Metropolis-adjusted Langevin algorithm (MALA) is a Metropolis-Hastings sampler that stems from the discretization of Langevin diffusion processes (see [13], for instance). Given the current state $\mathbf{X}[j] = \mathbf{x}$, the version with an isotropic step size generates candidates according to the proposal distribution

$$\mathbf{Y}[j+1] \sim \mathcal{N}\left(\mathbf{x} + \tfrac{h}{2}\nabla \log \pi(\mathbf{x}) \ , \ hI_n\right) \ , \tag{10}$$

for a chosen step size $h > 0$. The candidates are then accepted with probability $\alpha(\mathbf{x};\mathbf{y}) = 1 \wedge \pi(\mathbf{y})q_n(\mathbf{x};\mathbf{y})/\{\pi(\mathbf{x})q_n(\mathbf{y};\mathbf{x})\}$, where $q_n(\mathbf{y};\mathbf{x})$ is the density of the above normal proposal.

Since it uses local problem-specific information about the target, this sampler often yields results that are superior to isotropic RWM samplers, while frequently being almost as easy to implement. When the covariance matrix of the target components greatly departs from the identity matrix, one may instead use a pre-conditioned MALA :

$$\mathbf{Y}[j+1] \sim \mathcal{N}\left(\mathbf{x} + \tfrac{h}{2}A\nabla \log \pi(\mathbf{x}) \ , \ hA\right) \ , \tag{11}$$

where A is an $n \times n$ positive-definite matrix (see [14]).

More recently, [6] proposed a manifold variant of the Metropolis-adjusted Langevin algorithm (MMALA) based on the discretization of a diffusion with a position-dependent volatility matrix. This sampler generates candidates according to

$$\mathbf{Y}[j+1] \sim \mathcal{N}\left(\mathbf{x} + \tfrac{h}{2}G^{-1}(\mathbf{x})\nabla \log \pi(\mathbf{x}) + h\Omega(\mathbf{x}) \ , \ hG^{-1}(\mathbf{x})\right) \ ,$$

where $G(\mathbf{x})$ is some positive-definite $n \times n$ matrix, and

$$\Omega_i(\mathbf{x}) \quad = \quad |G(\mathbf{x})|^{-1/2} \sum_{j=1}^{n} \tfrac{\partial}{\partial X_j}\left[G^{-1}_{ij}(\mathbf{x}) \ |G(\mathbf{x})|^{1/2}\right] \ .$$

The choice of the metric tensor G is arbitrary; a popular choice that allows adaptation to the local curvature of the target π is the Fisher-Rao metric tensor, *i.e.* that $G(\mathbf{x})$ is based on the expected Fisher information (in a Bayesian context, where the expectation is computed with respect to the observations). This version of the MMALA has been shown to perform efficiently in a number of examples. As noted by the authors, the main overheads arising from the use of MMALA are the development of analytical expressions for the parameters of the normal proposal distribution, and the computational cost from updating these same parameters.

The position-dependent MALA (PMALA) proposed by [17] has been developed based on the discretization of a diffusion that slightly differs from that used by [6]. This results in a sampler that is at least as efficient as MMALA in theory, while being computationally cheaper. In the numerical examples, we shall thus use PMALA instead of MMALA; this sampler generates candidates according to

$$\mathbf{Y}[j+1] \sim \mathcal{N}\left(\mathbf{x} + \tfrac{h}{2}A(\mathbf{x})\nabla \log \pi(\mathbf{x}) + h\Gamma(\mathbf{x}) \ , \ hA(\mathbf{x})\right) \ ,$$

where $A(\mathbf{x}) = G^{-1}(\mathbf{x})$ and $\Gamma_i(\mathbf{x}) = \frac{1}{2}\sum_{j=1}^{n} \frac{\partial}{\partial X_j} A_{ij}(\mathbf{x})$.

Asymptotically optimal scaling results similar to those of [15] have been obtained for the isotropic MALA with product target density $\pi(\mathbf{x}^{(n)}) = \prod_{i=1}^{n} f(x_i)$. Of course, regularity assumptions stronger than those in Section 2 need to be imposed on the one-dimensional density f, such as finiteness of all moments $k \geq 1$, as well as polynomial bounds on the log-density and its derivatives; we refer the reader to [13] for more details.

In their article, [13] showed that for a step size of the form $h(n) = \ell^2/n^{1/3}$ with $\ell > 0$ the first component of an appropriately rescaled isotropic MALA algorithm asymptotically behaves according to a Langevin diffusion, where

$$v_{\text{MALA}}(\ell) = 2\ell^2\Phi\left(-\frac{\ell^3}{2}K\right) \equiv 2\ell^2\Phi\left(-\frac{\ell^3}{2}\sqrt{\mathbb{E}_X\left[\frac{5\{\frac{d^3}{dX^3}\log f(X)\}^2 - 3\{\frac{d^2}{dX^2}\log f(X)\}^3}{48}\right]}\right)$$

is the speed of the limiting diffusion (in which the positive root $K > 0$ is implicitly defined). The value ℓ optimizing the speed function satisfies

$$\hat{\ell}_{\text{MALA}} = \left(\frac{1.1236}{K}\right)^{1/3}$$

and yields an asymptotically optimal acceptance rate $a(\hat{\ell}_{\text{MALA}}) = 2\Phi(-1.1236/2) \approx 0.574$.

Of course, any of the described variants may be used in updating blocks of Metropolis-within-Gibbs samplers. Although we do not chase the details of such a proof, we expect optimal scaling results similar to those of Section 4 to hold for MALA-within-Gibbs algorithms. For the target density in (1), denote $\mathbf{x}_{(p+1):(p+n)} = \mathbf{x}$. For MALA-within-Gibbs, the drift term for the block update should be based on the conditional density given the fixed parameters, $\mathbf{x}^*_{1:p}$, as follows

$$\mathbf{Y}[j+1] \sim \mathcal{N}\left(\mathbf{x} + \tfrac{h}{2}\nabla \log \pi(\mathbf{x}|\mathbf{x}^*_{1:p})\ ,\ hI_n\right)\ .$$

Just as it makes sense to adjust the drift of the proposal distribution according to the most recent position of the mixing components, it would equally make sense to take this information into account when adjusting the proposal variance. Specifically, we define

$$\hat{\ell}_{\text{MALA}}(\mathbf{x}^*_{1:p}) = \frac{(1.1236)^{1/3}}{\{K(\mathbf{x}^*_{1:p})\}^{1/3}}\ , \tag{12}$$

with $K^2(\mathbf{x}^*_{1:p}) = (5\mathbb{E}_X[\{\frac{\partial^3}{\partial X^3}\log f(X|\mathbf{x}^*_{1:p})\}^2] - 3\mathbb{E}_X[\{\frac{\partial^2}{\partial X^2}\log f(X|\mathbf{x}^*_{1:p})\}^3])/48$; we then propose to use local step sizes $\hat{h}(n, \mathbf{x}^*_{1:p}) = \hat{\ell}^2_{\text{MALA}}(\mathbf{x}^*_{1:p})/n^{1/3}$, which correspond to an asymptotically optimal acceptance rate of 0.574. A fixed step size would also be available by optimizing the expected speed function $\mathbb{E}[v_{\text{MALA}}(\ell, \mathbf{X}_{1:p})] = \mathbb{E}[2\ell^2\Phi(-\ell^3 K(\mathbf{X}_{1:p})/2)]$ with respect to ℓ. For a target density $f(x|\mathbf{x}_{1:p})$ that is smooth enough with respect to x and $\mathbf{x}_{1:p}$, we expect the above local and fixed step sizes to be asymptotically optimal. The proof is beyond the scope of this paper, but we explore the idea in numerical studies.

The MALA-within-Gibbs with local step sizes shares some similarities with PMALA- and MMALA-within-Gibbs in which the first p blocks are updated as before (usual MALA), and a PMALA or MMALA is used to update the last block. In all three cases, step sizes vary across iterations; in contrast to PMALA and MMALA however, local step sizes in the MALA-within-Gibbs are independent of the current state $\mathbf{x}_{(p+1):(p+n)}$.

We could thus view local step sizes in the MALA-within-Gibbs as arising from a special diagonal metric tensor, whose diagonal entries are the expectation of a function (involving the second and third derivatives of the log-density) with respect to the parameters $\mathbf{X}_{(p+1):(p+n)}$ given $\mathbf{X}_{1:p}$. This expectation leads to the elimination of the metric tensor dependency on $\mathbf{x}_{(p+1):(p+n)}$. Note that, if desired, an extra expectation with respect to potentially present observations could also be computed, as in [6]. Since local step sizes depend on $\mathbf{x}^*_{1:p}$ only, which are updated separately, then $\Omega_i(\mathbf{x}) = \Gamma_i(\mathbf{x}) = 0$, $i = p+1, \ldots, p+n$ and the drift indeed remains the same as before.

The exact expression selected for the metric tensor is expected to naturally arise from weak convergence results, and leads to a simple and numerically cheap version of the PMALA- and MMALA-within-Gibbs (as it is independent of $\mathbf{x}_{(p+1):(p+n)}$). In addition, this choice of metric tensor conveniently comes with a guideline for an appropriate choice of ℓ. Of course, full-dimensional PMALA and MMALA are generally not equivalent to the MALA-within-Gibbs with local step sizes, as correlation between components involves a non-diagonal proposal covariance matrix for those samplers.

6. Hierarchies through location and scale

The tuning results of Sections 4 and 5 may be refined when $p = 1$ in (1) and the mixing parameter X_1 acts through the location or scale of the n conditionally i.i.d. components (or if X_1, X_2 are location and scale parameters in a target as in (1) with $p = 2$).

In this section, we study a hierarchical model with no observation; in this case, all the calculations required to derive local proposal variances are tractable. In practice however, observations would be present and local proposal variances could still be obtained analytically, as long as the conditional distribution of $\mathbf{X}_{(p+1):(p+n)}$ given $\mathbf{X}_{1:p}$ and the observations is tractable. This would be the case, for instance, if the distribution of $\mathbf{X}_{(p+1):(p+n)}|\mathbf{X}_{1:p}$ is conjugate to the conditional distribution of the observations given $\mathbf{X}_{(p+1):(p+n)}$.

6.1. Location mixing parameter

Consider a target density π as in (1), with $p = 1$ and $f(x_i|x_1) = f(x_i - x_1)$ for $i = 2, \dots, n+1$. The component X_1 thus acts as the (random) location parameter of the components $X_2, \dots, X_{n+1}$. In this particular context, we may use a simple change of variable ($u = x - x_1^*$) to reepxress $\mathcal{I}(x_1^*)$ in (7)

$$\mathbb{E}_X\left[\left(\tfrac{\partial}{\partial X}\log f(X - x_1^*)\right)^2\right] \;=\; \int_{\mathbb{R}}\left(\tfrac{\partial}{\partial u}\log f(u)\right)^2 f(u)\, du\ .$$

The locally optimal proposal variance of the RWM-within-Gibbs sampler is thus independent of x_1^*, *i.e.* $\hat{\ell}^{loc} = \hat{\ell}(0)$, where the function $\hat{\ell}(x_1)$ is given by (7). This is intuitively clear, as the smoothness of the target distribution is not affected by the location parameter.

Applying similar changes of variables, it is easy to show that the value of the local step size in the MALA-within-Gibbs also is independent of x_1^*, *i.e.* $\hat{\ell}^{loc}_{\text{MALA}} = \hat{\ell}_{\text{MALA}}(0)$, where the function $\hat{\ell}_{\text{MALA}}(x_1)$ is given by (12).

6.2. Scale mixing parameter

Now consider a target density as in (1), with $p = 1$ and $f(x_i|x_1) = \frac{1}{x_1}f(\frac{x_i}{x_1})$, $i = 2, \dots, n+1$; in other words, the component X_1 acts as the (random) scale parameter of the variables $X_2, \dots, X_{n+1}$. Applying a change of variable ($u = x/x_1^*$), the term $\mathcal{I}(x_1^*)$ in (7) may now be expressed as

$$\mathbb{E}\left[\left(\tfrac{\partial}{\partial X}\log \tfrac{1}{x_1^*}f\left(\tfrac{X}{x_1^*}\right)\right)^2\right] \;=\; \left(\tfrac{1}{x_1^*}\right)^2\int\left(\frac{\frac{\partial}{\partial u}f\,(u)}{f\,(u)}\right)^2 f\,(u)\,du\ .$$

The locally optimal proposal variance of the RWM-within-Gibbs sampler thus takes a form that is similar to that found in [15], except that it is multiplied by $(x_1^*)^2$; indeed, $\hat{\ell}^{sc}(x_1^*) = \hat{\ell}(1)x_1^*$, where the function $\hat{\ell}(x_1)$ is given by (7). The larger is the scale parameter X_1 at a given time, the larger is the proposal variance used to generate candidates for $X_2, \dots, X_{n+1}$ within the same iteration.

We may apply similar changes of variables to the local step sizes of the MALA-within-Gibbs. We find $\hat{\ell}^{sc}_{\mathrm{MALA}}(x_1^*) = \hat{\ell}_{\mathrm{MALA}}(1)x_1^*$, where the function $\hat{\ell}_{\mathrm{MALA}}(x_1)$ is given by (12); local step sizes are thus multiplied by $(x_1^*)^2$. For a scale mixing parameter, the MALA-within-Gibbs with local step sizes is thus equivalent to a MMALA-within-Gibbs in which the first block is updated according to the usual MALA, and the metric tensor in the last block is chosen to be $-\mathbb{E}[H(\mathbf{X}_{2:(n+1)}|x_1^*)]$, where $H(\mathbf{x}_{2:(n+1)}|x_1^*)$ is the Hessian of the log-target for $\mathbf{X}_{2:(n+1)}|x_1^*$ and the expectation is with respect to the density $\prod_{i=2}^{n+1} f(x_i|x_1^*)$. Note that this metric tensor is pre-tuned according to anticipated asymptotic results.

It is worth mentioning that in the presence of observations, local proposal variances would not necessarily be proportional to the scale parameter X_1. They would rather be proportional to the scaling parameter of the conditional distribution of $\mathbf{X}_{2:(n+1)}$ given X_1 and the observations, which may itself be a function of X_1. We now combine these notions into a numerical example that considers two mixing components, *i.e.* that includes location and scale parameters.

6.3. Simulation study: Normal-gamma-Student hierarchical distribution

Consider an n-dimensional hierarchical target such that $X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \Gamma(3,1)$, and $X_i|\mathbf{X}_{1:2} \sim t_\nu(X_1, 1/\sqrt{X_2})$ for $i = 3,\dots,n$ with $\nu = 7$; the density of a non-standard Student-$t(\mu,\eta)$ 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 and the Student distribution destroys conjugacy.

To illustrate the theoretical results, we consider a 42-dimensional target and apply various algorithms to obtain samples from this distribution: single-block RWM and MALA (each with isotropic and diagonal tunings), RWM- and MALA-within-Gibbs (with fixed and local tunings), pre-conditioned MALA, position-dependent MALA, and the Adaptive Metropolis of [7].

Isotropic, single-block RWM. The components are updated according to a $\mathcal{N}(\mathbf{x}, \ell^2 I_n/n)$ proposal distribution. Although the target distribution falls slightly outside the framework specified in [2], we may still approximate the optimal proposal variance by optimizing the speed measure in (3). With two mixing parameters, we however rely on $\gamma(\mathbf{x}_{1:2}, \mathbf{z}_{1:2}) = 0.8(1+z_1^2)x_2 + 0.35(z_2/x_2)^2$, where $\mathbf{z}_{1:2}$ come from independent $\mathcal{N}(0,1)$ random variables (see the discussion in [2]); note that this is independent of x_1. Letting $\mathbf{Z}_{1:2} \sim \mathcal{N}(0, I_2)$ and maximizing

$$\mathbb{E}_{\mathbf{X}_{1:2}}\left[\upsilon(\ell, \mathbf{X}_{1:2})\right] = 2\ell^2 \mathbb{E}_{\mathbf{X}_{1:2},\mathbf{Z}_{1:2}}\left[\Phi\left(-\frac{\ell}{2}\gamma^{1/2}(\mathbf{X}_{1:2}, \mathbf{Z}_{1:2})\right)\right]$$

with respect to ℓ leads to $\hat{\ell}_{\mathrm{R}} = 1.485$. This corresponds to an expected speed measure of $\mathbb{E}[\upsilon(\hat{\ell}_{\mathrm{R}}, \mathbf{X}_{1:2})] = 0.395$. Using the relationship between the speed function and expected acceptance rate, $\mathbb{E}[\upsilon(\ell, \mathbf{X}_{1:2})/\ell^2] = a(\ell)$ (see (8) and (9) for the RWM-within-Gibbs, but this is valid more generally), a corresponding acceptance rate of $\mathbb{E}[\upsilon(\hat{\ell}_{\mathrm{R}}, \mathbf{X}_{1:2})]/\hat{\ell}^2_{\mathrm{R}} = 0.179$ is obtained.

Diagonal, single-block RWM. The proposal variances of the RWM sampler are the marginal variances of the target model, multiplied by the tunable factor ℓ^2/n; we choose ℓ to yield an acceptance rate between 0.179 and 0.234.

Local RWM-within-Gibbs. We update X_1, X_2, and $\mathbf{X}_{3:n}$ separately, using normal proposal distributions. The proposal standard deviations of the first and second blocks are set to $\sigma_1 = 0.25$ and $\sigma_2 = 1.5$ respectively, leading to acceptance rates close to 45% for the corresponding sub-algorithms. In practice, a convenient approach is to fix the proposal variance of the block $\mathbf{X}_{3:n}$, and then tune σ_1, σ_2 to attain the desired acceptance rates.

Hereafter, $\mathbf{x}^*_{1:2}$ and $\mathbf{x}_{3:n}$ refer to the latest available states of $\mathbf{X}_{1:2}$ and $\mathbf{X}_{3:n}$, just after $\mathbf{X}_{1:2}$ is updated. For a Student-t distribution, we find

$$\mathcal{I}(\mathbf{x}^*_{1:2}) = x_2^* \, \frac{(\nu+1)^2}{\nu(\nu+2)} \, \frac{\Gamma((\nu+1)/2)\,\Gamma((\nu+4)/2)}{\Gamma(\nu/2)\,\Gamma((\nu+5)/2)} \;=\; 0.8\; x_2^* \;;$$

the locally optimal variance arising from (7) is thus $\hat{\ell}^2_{\mathrm{M}}(\mathbf{x}^*_{1:2})/(n-2) = (2.38)^2/\{0.8x^*_2(n-2)\}$. When X_2 is large (small), the distribution of $\mathbf{X}_{3:n}$ is narrowed down (spread out) and accordingly, the proposal variance must be small (large). The expected efficiency based on these local tunings is $\mathbb{E}\left[2\hat{\ell}^2_{\mathrm{M}}(\mathbf{X}_{1:2})\Phi\left(-2.38/2\right)\right] = 0.828$, which of course corresponds to an acceptance rate of 0.234.

Fixed RWM-within-Gibbs. The proposal variance of the third block is now fixed, and maximizes $\mathbb{E}\left[2\ell^2\Phi\left(-\ell\sqrt{0.8X_2}/2\right)\right]$. The resulting variance is $\hat{\ell}^2_{\mathrm{M}}/(n-2) = (1.90)^2/(n-2)$ and corresponds to asymptotic acceptance rate and expected efficiency of 0.191 and 0.691 respectively. There is thus an improvement of 20% available from tuning the proposal variance locally, as opposed to favoring a fixed approach for the RWM-within-Gibbs.

Isotropic, single-block MALA. All n components are simultaneously updated according to (10). In the case of hierarchical target distributions as in (1), we are not aware of optimal scaling results for tuning MALA; we thus settle for an acceptance rate of 0.35, which is close to optimal according to the efficiency curves in Figure 2 (to be described shortly).

Diagonal, single-block MALA. The step sizes of the MALA sampler are the marginal variances of the target model, multiplied by the tunable factor $\ell^2/n^{1/3}$; we choose ℓ to yield an acceptance rate between 0.35 and 0.57.

Pre-conditioned MALA. All n components are simultaneously updated according to (11). We set $A = -\mathbb{E}[H(\mathbf{X}_{1:n})]^{-1}$, where $H(\mathbf{x}_{1:n})$ is the Hessian of the log-target and the expectation is with respect to the target distribution, and tune the step size so as to yield an acceptance rate close to 0.574.

Local MALA-within-Gibbs. We update X_1, X_2, and $\mathbf{X}_{3:n}$ separately, using MALA samplers. The step sizes of the first two blocks are set to $h_1 = 0.2$ and $h_2 = 1.1$ respectively, leading to acceptance rates close to 57% for the corresponding sub-algorithms. As discussed in Section 3 of [13], this is a conservative choice in low-dimensional settings.

For a Student-t distribution, we find $K(\mathbf{x}^*_{1:2}) = 0.262(x^*_2)^{3/2}$ and from (12), this leads to locally optimal step sizes $\hat{h}(n, \mathbf{x}^*_{1:2}) = 2.64/\{x^*_2(n-2)^{1/3}\}$. The expected efficiency based on these local tunings is $\mathbb{E}\left[2\hat{\ell}^2_{\mathrm{MALA}}(\mathbf{X}_{1:2})\Phi\left(-1.1236/2\right)\right] = 0.761$, which of course corresponds to an acceptance rate of 0.574.

Fixed MALA-within-Gibbs. The step size of the third block is now fixed, and maximizes $\mathbb{E}\left[2\ell^2\Phi\left(-\ell^3 0.262(X_2)^{3/2}/2\right)\right]$. The resulting step size is $\hat{\ell}^2_{\mathrm{MALA}}/(n-2)^{1/3} = (1.07)^2/(n-2)^{1/3}$ and corresponds to asymptotic acceptance rate and expected efficiency of 0.467 and 0.535 respectively. There is thus an improvement of more than 40% available from tuning the step sizes locally, as opposed to favoring a fixed approach for the MALA-within-Gibbs.

Theoretical performances of MALA and RWM cannot be compared directly. To obtain weak convergence results for RWM and MALA, one needs to rescale space and time, using factors of n and $n^{1/3}$ respectively. For high-dimensional targets, the theoretical efficiency obtained through MALA's speed function should thus be multiplied by an $\mathcal{O}(n^{2/3})$ factor in order to be compared to the theoretical efficiency of a RWM. We then expect a significant gain from using a locally tuned MALA-within-Gibbs over its RWM-within-Gibbs counterpart. This efficiency gain is of course tempered by extra computations, required at each iteration of the sampler.

Position-dependent MALA. The metric tensor suggested in [6] is $G(\mathbf{x}_{1:n}) = -H(\mathbf{x}_{1:n})$, where $H(\mathbf{x}_{1:n})$ is the Hessian of the log-target. Since the Student-t distribution is not log-concave, $-H(\mathbf{x}_{1:n})$ is not positive definite everywhere, and so neither is $A(\mathbf{x}_{1:n}) = -H^{-1}(\mathbf{x}_{1:n})$. We thus take $G(\mathbf{x}_{1:2}) = A^{-1}(\mathbf{x}_{1:2}) = -\mathbb{E}[H(\mathbf{x}_{1:2}, \mathbf{X}_{3:n})]$, where the expectation is with respect to the conditional distribution of $\mathbf{X}_{3:n}|\mathbf{x}_{1:2}$. This leads to a relatively simple form for $A(\mathbf{x}_{1:2})$, which conveniently allows to compute the extra drift term $\Gamma(\mathbf{x}_{1:2})$. This results in a sampler that is similar to the locally updated MALA-within-Gibbs, but where all n components are updated simultaneously (and not independently). We tune the acceptance rate to the usual 0.574.

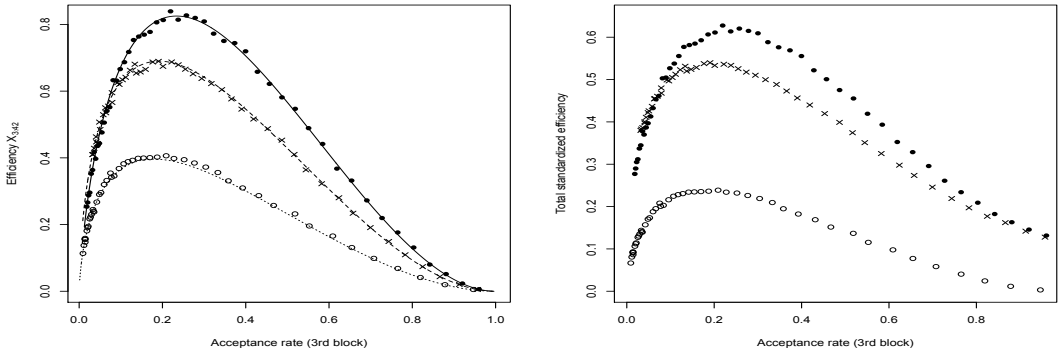

Figure 1: Efficiency against acceptance rate of RWM samplers for the normal-gamma-Student target. Left: efficiency based on $\mathbf{X}_{3:n}$. Right: total standardized efficiency based on $\mathbf{X}_{1:n}$. The solid, dashed, and dotted lines represent, in order, the theoretical efficiency curve of the local, fixed RWM-within-Gibbs, and isotropic RWM. Symbols come from simulations with $n = 42$ (top to bottom: local and fixed RWM-w-Gibbs, isotropic RWM).

Adaptive Metropolis. This sampler proposes candidates from a $\mathcal{N}(\mathbf{x}_{1:n}, \ell_{\mathrm{A}}^2 \Sigma_j / n)$, where Σ_j is an estimate of the target covariance matrix that is updated recursively at every iteration. Since the chain takes a long time to adapt, we shall use 1,000,000 iterations to update Σ, which will then remain fixed for the rest of the run. The initial covariance Σ_0 is a diagonal matrix whose entries are the marginal target variances. We rely on $\hat{\ell}_{\mathrm{A}} = 2.38$, as suggested in [7].

We now illustrate the performance of algorithms featuring a diagonal proposal covariance matrix with efficiency curves. Each of these samplers is run with 50 different tunings. Samplers that generate candidates from a multivariate normal distribution are expensive to run, so we consider optimal versions of these samplers later on. For each run, we perform 5,000,000 iterations and measure efficiency by recording the standardized expected squared jumping distance

$$\mathrm{ESJD} = \frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{n} \frac{1}{\omega_i^2} \left(x_i[j] - x_i[j-1] \right)^2 ; \tag{13}$$

here, N is the number of iterations, n is the dimension of the target distribution as before, and ω_i^2 is the marginal variance of the ith target component. We also record the average acceptance rate of each algorithm, expressed as

$$\mathrm{AAR} = \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}\{\mathbf{x}[j] \neq \mathbf{x}[j-1])\}.$$

The resulting curves of efficiency against acceptance rates are then compared in Figure 1 (isotropic RWM, fixed & local RWM-within-Gibbs) and in Figure 2 (isotropic MALA, fixed & local MALA-within-Gibbs). The left graph of each figure combines the efficiency curves of the 42-dimensional samplers along with their theoretical efficiency curves (expected speed measure against expected acceptance rate); it focuses on the efficiency measure arising from the third block only, *i.e.* ignoring, in (13), contributions coming from the movements of $\mathbf{X}_{1:2}$ and setting $\omega_i^2 = 1$. The right graph of each figure compares total standardized efficiency of the samplers.

Optimal versions of the samplers discussed are compared in Table 1, in which running times for performing 5,000,000 iterations are also reported (obtained with the function `system.time` in

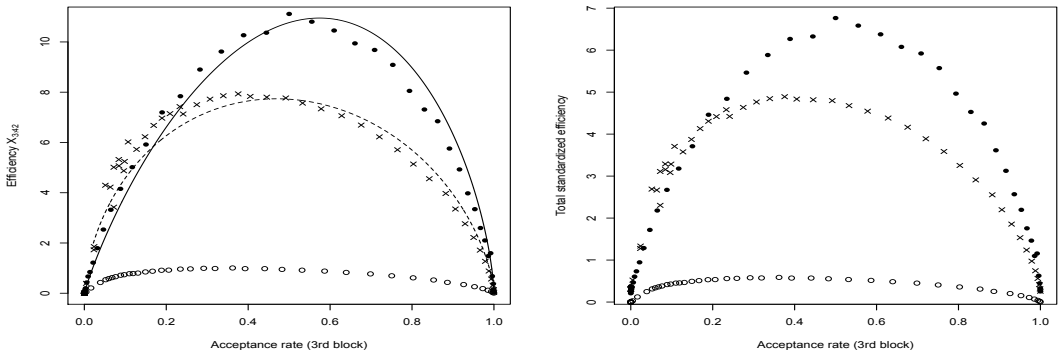

Figure 2: Efficiency against acceptance rate of MALA samplers for the normal-gamma-Student target. Left: efficiency based on $\mathbf{X}_{3:n}$. Right: total standardized efficiency based on $\mathbf{X}_{1:n}$. The solid and dashed lines respectively represent the theoretical efficiency curve of the local and fixed MALA-within-Gibbs. Symbols come from simulations with $n = 42$ (top to bottom: local and fixed MALA-within-Gibbs, and isotropic MALA).

R). We note that pre-conditioned MALA, position-dependent MALA, and Adaptive Metropolis samplers generate candidates from multivariate normal distributions with correlation, which is achieved by performing a Cholesky decomposition of the covariance matrix and generating n independent standard normal observations. Hereafter, net efficiency refers to an efficiency measure that penalizes for computational effort; it is defined as the ESJD divided by the running time.

According to Table 1, the pre-conditioned MALA yields the highest net efficiency: it offers a gain of 54% over the local version of MALA-within-Gibbs, which comes second. The latter offers a net efficiency gain of 35% over the fixed MALA-within-Gibbs, and is almost 3 times more efficient than the position-dependent MALA. Unsurprisingly, net efficiency of RWM samplers come behind those of the fanciest MALA algorithms. The position-dependent MALA is almost twice as efficient as the local version of the RWM-within-Gibbs, which constitutes the best available option among RWM-type samplers. In the current context, the latter results in an efficiency gain of about 10% compared to the diagonal, single-block MALA, 15% compared to the fixed RWM-within-Gibbs, and 30% compared to the Adaptive Metropolis sampler. Single-block RWM (isotropic or diagonal) and MALA (isotropic) do not constitute competitive options here, but rather act as baselines for the other methods. Although the pre-conditioned MALA offers a significant improvement compared to other competitors, we remind the reader that this target model is used as a proof of concept. A real example shall be studied in Section 8.

7. Inhomogeneous proposal variances

The theoretical results of Sections 4, 5, and 6 may be extended to target densities

$$\pi(\mathbf{x}^{(n+p)}) \quad = \quad f_1(\mathbf{x}_{1:p}) \prod_{i=p+1}^{p+n} \frac{1}{C_i} f\left(\frac{x_i - M_i}{C_i}\right) , \tag{14}$$

where f_1, f are densities satisfying the assumptions stated in Section 2, and $M_i \equiv M_i(\mathbf{x}_{1:p})$, $C_i \equiv C_i(\mathbf{x}_{1:p})$ $(i = p+1, \ldots, p+n)$ respectively are location and scale parameters which are conditionally i.i.d. (given $\mathbf{X}_{1:p}$) from a distribution with finite variance. In other words, the X_is

Table 1: Standardized efficiency, acceptance rate, and running times of optimized samplers applied to the normal-gamma-Student target.

Sampler	AAR	ESJD	Time (sec)	**Sampler**	AAR	ESJD	Time (sec)
Isotropic RWM	0.209	0.239	211	Isotropic MALA	0.362	0.587	455
Diagonal RWM	0.204	0.258	228	Diagonal MALA	0.442	0.882	518
Fixed RWM-w-G	0.191	0.540	330	Fixed MALA-w-G	0.376	4.891	621
Loc. RWM-w-G	0.215	0.628	338	Loc. MALA-w-G	0.500	6.765	632
Adapt. Met.	0.267	0.860	605	Prec. MALA	0.552	14.28	864
				PMALA	0.571	13.25	3 634

are conditionally independent given $\mathbf{X}_{1:p}$, and are distributed according to a common density f that features different location and scale parameters.

In a RWM-within-Gibbs, every block is updated with respect to its full conditional distribution. Denoting the updated values of the scales at time $t+1$ by $C_i^* \equiv C_i(\mathbf{x}_{1:p}^*)$ ($i = p+1, \ldots, p+n$), we generate a candidate for X_i according to a Gaussian proposal with variance of the form $\ell_i^2(\mathbf{x}_{1:p}^*)/n \equiv \tilde{\ell}^2 C_i^2(\mathbf{x}_{1:p}^*)/n$ for some constant $\tilde{\ell} > 0$.

Following [12], we then believe Corollary 3 in Section 4.1 to be still valid. In particular, we expect the process associated to the ith component to asymptotically behave according to the diffusion and speed measure in (5) and (6), but with the function $f(x_i|M_i(\mathbf{x}_{1:p}^*), C_i(\mathbf{x}_{1:p}^*))$ as defined in (14). This would give rise to a locally optimal proposal variance for the ith component that is equal to $\hat{\ell}_i^2(\mathbf{x}_{1:p}^*)/n = \hat{\ell}^2(0,1)C_i^2(\mathbf{x}_{1:p}^*)/n$, with the function $\hat{\ell}(M^*, C^*)$ as in (7). In particular, the asymptotically optimal acceptance rate would still be 0.234. Proofs of the asymptotic results found in this paper heavily rely on Laws of Large Numbers and Central Limit Theorems. It is thus unclear whether or not optimal scaling results based on inhomogeneous proposal variances could be generalized to target densities other than (14).

Similar extensions are also believed to hold for MALA-within-Gibbs samplers; in particular, candidates in the last block should be generated according to step sizes $\hat{h}_i(n, \mathbf{x}_{1:p}^*) = \hat{\ell}_{\text{MALA}}^2(0,1)C_i^2(\mathbf{x}_{1:p}^*)/n^{1/3}$ for $i = p+1, \ldots, p+n$, with $\hat{\ell}_{\text{MALA}}(M^*, C^*)$ as in (12). As mentioned previously, it is however unclear as to what regularity assumptions should be imposed on the target density for these extensions to hold.

To illustrate the correspondence between these heuristic claims and their anticipated asymptotic behaviours, consider the example of Section 6.3 in which there is no location parameter: $X_1 \sim \Gamma(3,1)$ and $X_i|X_1 \sim t_7(0, \xi_i/X_1)$, with fixed $\xi_i = i-1$ for $i = 2, \ldots, 41$. All else being equal, the 5,000,000-iteration simulation study described in that section for the local RWM- and MALA-within-Gibbs leads to the graphs in Figure 3. In these graphs, the RWM-within-Gibbs relies on the proposal variances $\ell_i^2(x_1^*)/n = \tilde{\ell}^2 \xi_i / \{n x_1^* \mathcal{I}_i(x_1^* = \xi_i)\}$, $i = p+1, \ldots, p+n$ (for 50 values of $\tilde{\ell}$ equally spaced in $[0.1, 5]$); the MALA-within-Gibbs relies on step sizes $\ell_{\text{MALA},i}^2(x_1^*)/n^{1/3} = \tilde{\ell}^2 \xi_i / \{x_1^* K_i^{2/3}(x_1^* = \xi_i) n^{1/3}\}$, $i = p+1, \ldots, p+n$ (for 50 values of $\tilde{\ell}$ equally spaced in $[0.05, 2.5]$). Simulations for the local RWM-within-Gibbs approach are compared to the standardized theoretical efficiency curve of

$$\frac{1}{40}\sum_{i=2}^{41} \frac{\mathbb{E}[\upsilon_i(\ell, X_1)]}{\xi_i} = \frac{1}{40}\sum_{i=2}^{41} \frac{2\tilde{\ell}^2 \mathbb{E}[X_1^{-1}]}{\mathcal{I}_i(X_1 = \xi_i)} \Phi\left(-\frac{\tilde{\ell}}{2}\right)$$

versus the expected acceptance rate $2\Phi(-\tilde{\ell}/2)$. For MALA-within-Gibbs, standardized theoret-

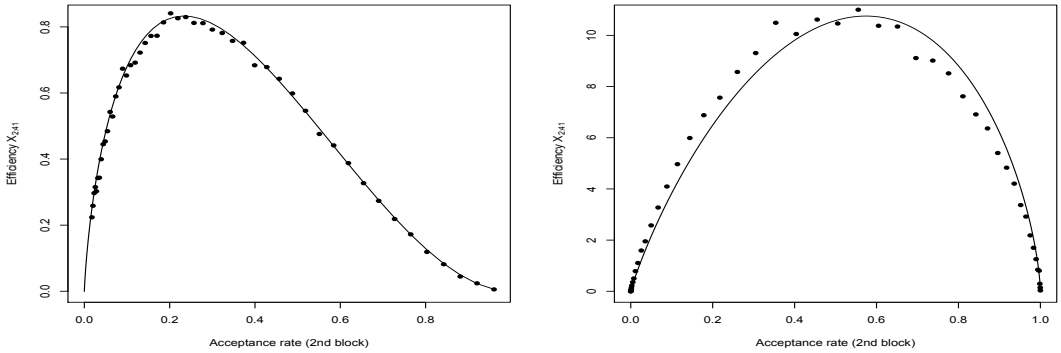

Figure 3: Standardized efficiency of $\mathbf{X}_{2:n}$ against acceptance rate for the gamma-Student target. Left : RWM-within-Gibbs. Right : MALA-within-Gibbs. The solid lines represent the theoretical standardized efficiency curves; the symbols come from simulations with $n = 41$.

ical efficiency satisfies

$$\frac{1}{40}\sum_{i=2}^{41}\frac{\mathbb{E}[v_i^{\mathrm{MALA}}(\ell, X_1)]}{\xi_i} = \frac{1}{40}\sum_{i=2}^{41}\frac{2\tilde{\ell}^2\mathbb{E}[X_1^{-1}]}{\{K_i(X_1=\xi_i)\}^{2/3}}\Phi\left(-\frac{\tilde{\ell}^3}{2}\right) ,$$

which is plotted against the expected acceptance rate $2\Phi(-\tilde{\ell}^3/2)$.

Under this inhomogeneous setting, the relative performances and computational efforts of the samplers previously studied are as discussed in Section 6.3. Of course, if proposal variances of samplers such as RWM, MALA, fixed RWM- and MALA-within-Gibbs are not adjusted to take inhomogeneity into account, the gap between their performance and that of the other samplers is expected to significantly widen.

8. A stochastic volatility model

Consider a stochastic volatility model in the flavour of [9] and [6], in which the latent volatilities take the form of an autoregressive process of order 1. That is, $D_i = \varepsilon_i \exp\{X_i/2\}$ with $X_{i+1} = \phi X_i + \eta_{i+1}$, where $\varepsilon_i \sim \mathcal{N}(0,1)$, $\eta_i \sim \mathcal{N}(0,\tau^2)$ and $X_1 \sim \mathcal{N}(0,\tau^2/(1-\phi^2))$. Priors for the parameters are $\tau^2 \sim \mathrm{I}\Gamma(\delta,\lambda)$ and $(\phi+1)/2 \sim \beta(a,b)$, where $\mathrm{I}\Gamma(\delta,\lambda)$ is the inverse gamma distribution with density proportional to $x^{-(\delta+1)}\mathrm{e}^{-\lambda x}$. This model leads to an $(n+2)$-dimensional posterior density $\pi(\tau^2,\phi,X_1,\ldots,X_n|\mathbf{d}_{1:n})$.

The posterior density is too complex for analytic computation, and numerical integration must be ruled out due to the high-dimensionality of the problem. This distribution is best sampled with MCMC methods; in the current setting, we propose to use RWM- and MALA-within-Gibbs samplers with three blocks of variables, τ^2, ϕ, and $\boldsymbol{X}_{1:n}$. We are also interested in assessing the performance of competitors such as those considered in Section 6.3.

In Section 6.3, it was not necessary to apply a change of variable to take care of the positivity constraint on the scale parameter X_2. Indeed, the target density was smooth on $\mathbb{R}$, in the sense that the density was converging relatively slowly to 0 as X_2 approached 0. The hyperparameters used in the current example lead to a target density that is not as smooth, so we let $\tau^2 = \exp\{\kappa\}$

and $\phi = \tanh(\gamma)$; this yields the following $(n+2)$-dimensional posterior density

$$\pi(\kappa,\gamma,\mathbf{x}_{1:n}|\mathbf{d}_{1:n}) \propto \exp\left\{-\kappa(\tfrac{n}{2}+\delta)\right\} \frac{\mathrm{e}^{-\gamma(2b+1)}}{(1+\mathrm{e}^{-2\gamma})^{a+b+1}} \exp\left\{-\tfrac{1}{2}\sum_{i=1}^{n}(x_i + d_i^2 \mathrm{e}^{-x_i})\right\}$$
$$\times \exp\left\{\frac{\mathrm{e}^{-\kappa}}{2}\left[\lambda + \frac{4\mathrm{e}^{-2\gamma}}{(1+\mathrm{e}^{-2\gamma})^2}x_1^2 + \sum_{i=2}^{n}\left(x_i - (\tfrac{1-\mathrm{e}^{-2\gamma}}{1+\mathrm{e}^{-2\gamma}})x_{i-1}\right)^2\right]\right\} .$$

In the following experimental study, we fix the hyperparameters to $\delta = 1$, $\lambda = 1.5$, $a = 20$, and $b = 6$, and consider a 100-dimensional dataset $\mathbf{d}_{1:100}$ in which the data points exhibit low correlation. This dataset is simulated directly from the target model with $\phi = 0.15$ and $\tau^2 = 0.75$. Of interest is to evaluate the performance of previously described approaches, in the context of a target model falling outside the assumptions of the theory.

We perform a 5,000,000-iteration simulation study similar to those described previously. The ESJD is again used as a measure of efficiency; it is reported in Table 2 for each sampler, along with acceptance rates and running times. We now point out the particularities of the algorithms implemented.

Adaptive Metropolis. The initial proposal covariance matrix Σ_0 is a diagonal matrix of dimension $n+2$. The diagonal elements are the reciprocal of those in $-\mathbb{E}[H(\kappa,\gamma,\mathbf{X}_{1:n}|\mathbf{D}_{1:n})]$, where H is the Hessian of the log-posterior density and the expectation is computed with respect to the distribution of $(\kappa,\gamma,\mathbf{X}_{1:n},\mathbf{D}_{1:n})$. For instance, the diagonal entry related to κ is $-\mathbb{E}[(\frac{\partial^2}{\partial\kappa^2}\log\pi(\kappa,\gamma,\mathbf{X}_{1:n}|\mathbf{D}_{1:n}))^2]^{-1} = \delta + n/2$. Since the adaptation of Σ_j is slow, we again use 1,000,000 iterations to update Σ, which then remains fixed for the rest of the run. We tune the acceptance rate as close as possible to 0.234.

Pre-conditioned MALA. The matrix A in (11) satisfies $A^{-1} = -\mathbb{E}[H(\kappa,\gamma,\mathbf{X}_{1:n}|\mathbf{D}_{1:n})]$, where the expectation is with respect to the distribution of $(\kappa,\gamma,\mathbf{X}_{1:n},\mathbf{D}_{1:n})$. The matrix A^{-1} is tridiagonal; for $i = 1,\ldots,n-1$ we have, for instance, $A^{-1}_{i+3,i+2} = A^{-1}_{i+2,i+3} = -(\frac{a-b}{a+b})\frac{\delta}{\lambda}$. We settle for the usual acceptance rate of 0.574.

Local RWM-within-Gibbs. We update κ, γ, and $\mathbf{X}_{1:n}$ separately; the proposal standard deviations of the first and second blocks are set to 0.2 and 0.27 respectively, leading to acceptance rates close to 45% for the corresponding sub-algorithms. Tuning of the different blocks were conveniently performed simultaneously and independently, as the scaling of a particular block did not affect the tuning of the other blocks.

Hereafter, κ^*, γ^* and $\mathbf{x}_{1:n}$ refer to the latest available states of the process, just after κ and γ are updated. An n-dimensional update of $\mathbf{X}_{1:n}$ in the third block is performed with respect to the conditional density $\pi(\mathbf{x}_{1:n}|\kappa^*,\gamma^*,\mathbf{d}_{1:n})$. When working conditionally on data points $\mathbf{d}_{1:n}$, the expectation in $\mathcal{I}_i(\kappa^*,\gamma^*) = \mathbb{E}[(\frac{\partial}{\partial X_i}\log\pi(\mathbf{X}_{1:n}|\kappa^*,\gamma^*,\mathbf{d}_{1:n}))^2]$ $(i = 1,\ldots,n)$ is not necessarily easy to obtain, as the full conditional distribution is usually not as simple as initial distributions. This problem may be solved by first computing the above expectation with respect to the random variables D_i $(i = 1,\ldots,n)$, and then with respect to $\mathbf{X}_{1:n}$. This yields local proposal variances that have been averaged over all possible datasets, *i.e.* the best possible local proposal variances without information about the specific dataset at hand. This approach has also been favored in [6], in the choice of their metric tensor. The only cases where this extra expectation would possibly result in poor proposal variances would be when the dataset is highly improbable given the target model. In the current situation, local proposal variances for the third block are

$$\frac{\ell^2}{n}\left(\tfrac{1}{2}+\mathrm{e}^{-\kappa^*}, \tfrac{1}{2}+\mathrm{e}^{-\kappa^*}\left(1+(\tfrac{1-\mathrm{e}^{-2\gamma^*}}{1+\mathrm{e}^{-2\gamma^*}})^2\right), \ldots, \tfrac{1}{2}+\mathrm{e}^{-\kappa^*}\left(1+(\tfrac{1-\mathrm{e}^{-2\gamma^*}}{1+\mathrm{e}^{-2\gamma^*}})^2\right), \tfrac{1}{2}+\mathrm{e}^{-\kappa^*}\right)^{-1} .$$

With these proposal variances, we simply tune the acceptance rate as close as possible to 0.234.

Fixed RWM-within-Gibbs. The theory of Section 4.2 is used to obtain an approximately optimal acceptance rate of 0.2, keeping in mind that regularity assumptions are of course violated here (and that acceptance rates generally seem more robust than proposal variances, at least for RWM samplers).

Local MALA-within-Gibbs. We update κ, γ, and $\mathbf{X}_{1:n}$ separately, using MALA samplers. The step sizes of the first and second blocks are set to $h_1 = 0.2$ and $h_2 = 0.27$ respectively, leading to acceptance rates close to 57% for the corresponding sub-algorithms. As for the RWM-within-Gibbs sampler, we obtain local step sizes by computing an extra expectation with respect to $\mathbf{D}_{1:n}$. This leads to step sizes $\frac{\ell^2}{n^{1/3}}(\frac{1}{K_1}, \frac{1}{K_2}, \ldots, \frac{1}{K_n})^{2/3}$, where

$$\begin{aligned} K_1^2 &= K_n^2 = \frac{1}{128}(25 + 18\mathrm{e}^{-\kappa^*} + 12\mathrm{e}^{-2\kappa^*} + 8\mathrm{e}^{-3\kappa^*}) , \\ K_i^2 &= \frac{1}{128}\left[25 + 18\mathrm{e}^{-\kappa^*}\left(1 + (\tfrac{1-\mathrm{e}^{-2\gamma^*}}{1+\mathrm{e}^{-2\gamma^*}})^2\right) + 12\mathrm{e}^{-2\kappa^*}\left(1 + (\tfrac{1-\mathrm{e}^{-2\gamma^*}}{1+\mathrm{e}^{-2\gamma^*}})^2\right)^2 \right. \\ &\qquad \left. + 8\mathrm{e}^{-3\kappa^*}\left(1 + (\tfrac{1-\mathrm{e}^{-2\gamma^*}}{1+\mathrm{e}^{-2\gamma^*}})^2\right)^3\right] , \quad i = 2, \ldots, n-1 . \end{aligned}$$

We then choose ℓ so that the acceptance rate be close to 0.574.

Fixed MALA-within-Gibbs. The sampler is tuned according to the approximated optimal acceptance rate of 0.54 (numerically computed, using concepts in Section 5).

Position-dependent MALA. Due to the difficulty in computing the derivatives of the inverse metric tensor, we instead do as in [6] and use a position-dependent MALA to update the last block of a MALA-within-Gibbs. We rely here on the same three blocks as before: the first two blocks are tuned as detailed above, while the last block relies on the metric tensor $A^{-1}(\kappa^*, \gamma^*, \mathbf{X}_{1:n}) = -\mathbb{E}[H(\mathbf{X}_{1:n}|\kappa^*, \gamma^*, \mathbf{D}_{1:n})]$, where H is the Hessian of the log of the density $\pi(\mathbf{x}_{1:n}|\kappa^*, \gamma^*, \mathbf{d}_{1:n})$ and the expectation is with respect to the distribution of $\mathbf{D}_{1:n}$ only. The resulting matrix turns out to be tridiagonal, with entries that are independent of $\mathbf{X}_{1:n}$. Accordingly, $\Gamma_i(\mathbf{x}) = 0$, $i = 1, \ldots, n$, and the drift term is the same as in a classical MALA. This is thus equivalent to using a pre-conditioned MALA for the third block (we refer the reader to [6] for more details). We tune the acceptance rate to the usual 0.574.

Table 2 is similar in form to Table 1 and compares the ESJD of each sampler along with their running times; the acceptance rates are also included. We set $\omega_i^2 = 1$ for $i = 1, \ldots, n$ in (13), as all the samplers implemented feature different proposal scalings for κ, γ, and $\mathbf{X}_{1:n}$, and can thus be compared directly.

The ESJD obtained behave as expected, showing a progression from RWM-type samplers to MALA-within-Gibbs, and eventually the fancier MALA with correlated candidates. Local versions of the RWM- and MALA-within-Gibbs lead to small gains of about 7% and 13% respectively, in terms of ESJD over their fixed counterparts. The position-dependent MALA is the sampler that results in the highest ESJD, while the fixed RWM-within-Gibbs and Adaptive Metropolis fight for the last place.

When including computational effort in the picture, the outcome is different. The need of generating correlated candidates in some samplers disturbs the above ordering. The local MALA-within-Gibbs seems to offer the best compromise, although its edge over the fixed version is tempered by a slightly increased computational effort. The pre-conditioned MALA suffers a drop of 40% in terms of net efficiency compared to the local MALA-within-Gibbs. In the current context, the computational overhead of the position-dependent MALA prevents this sampler from being a serious competitor to MALA-within-Gibbs samplers; it also suffers a net efficiency loss of 8% compared to the local RWM-within-Gibbs algorithm. The local and fixed RWM-within-Gibbs are virtually equivalent when accounting for computational effort, and are 3 times as efficient as the Adaptive Metropolis. The light variability among local proposal

Table 2: Efficiency, acceptance rate, and running times of optimized samplers applied to the stochastic volatility model.

Sampler	AAR	ESJD	Time (sec)	**Sampler**	AAR	ESJD	Time (sec)
Fixed RWM-w-G	0.169	0.527	499	Fixed MALA-w-G	0.521	11.174	1 089
Loc. RWM-w-G	0.239	0.564	531	Loc. MALA-w-G	0.559	12.627	1 184
Adapt. Met.	0.296	0.526	1 534	Prec. MALA	0.510	13.940	1 835
				PMALA	0.578	16.028	16 309

variances in a given iteration, and also from one iteration to the other, explains in part the modest improvement offered by local versions over fixed ones in the current context.

9. Discussion

Optimality results for the RWM-within-Gibbs sampler have been presented. It has been concluded that, compared to fixed proposal variances, local ones generally offer interesting benefits in terms of performance when sampling from hierarchical models involving scaling parameters. These variances are easily implemented, at a marginal additional computational cost compared to RWM-within-Gibbs with fixed variances. They also allow to rely on simple, well-known optimality results ($\hat{\ell} = 2.38/\{\mathcal{I}(\mathbf{x}^*_{1:p})\}^{1/2}$, leading to an acceptance rate of 0.234). The analytical derivation of local proposal variances however requires that the distribution of the conditionally independent components given the mixing parameters and the observations be tractable. When this is not the case, a fixed approach that implies numerically solving for optimal proposal variance and acceptance rate prior to running the sampler may be favored. It has been demonstrated that the optimal acceptance rate arising from a fixed proposal variance is no greater than 0.234. Similar conclusions have then be drawn for fixed versus local MALA-within-Gibbs samplers. When available, users should then favor local approaches over their fixed counterpart as they represent a safe yet interesting avenue, both for the RWM- and MALA-within-Gibbs.

An expression for locally optimal, inhomogeneous proposal variances has been proposed for the case where the full conditional densities of the n-dimensional block are inhomogeneous but belong to a location/scale family. The simulation study of Section 8 leads us to believe that local, inhomogeneous proposal variances and step sizes could be used more generally. Indeed, each proposal variance (step size) is adjusted according to the roughness of the target distribution in a particular direction, and so they constitute an intuitive option, even when the target density does not fit in the prescribed framework.

In the examples, the locally tuned MALA-within-Gibbs yielded convincing results, while net performances of the local RWM-within-Gibbs outdid those of RWM-type samplers (including the Adaptive Metropolis), as well as some of the MALA-type samplers. We obviously do not expect local MALA-within-Gibbs to outdo all competitors in all situations. Gibbs samplers, for instance, become very inefficient in the presence of strong correlation between blocks; the proposed approach does not circumvent this feature and is thus not expected to be competitive in strongly correlated situations. What we do believe, however, is that once a user has chosen to implement RWM- or MALA-within-Gibbs, he will do better by relying on a local version of these samplers. As witnessed from the simulations studies, the efficiency gain available from implementing a local RWM- or MALA-within-Gibbs over its fixed version is however largely influenced by the variability present in the hierarchical model, with a larger variability sustaining the need for local samplers.

Acknowledgements

This work has been supported by the Natural Sciences and Engineering Research Council of Canada. The author wishes to thank the anonymous referees for very constructive comments.

References

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

[2] Bédard M (2015) Hierarchical models and the tuning of RWM algorithms, submitted

[3] Bédard M, Douc R, Moulines E (2014) Scaling analysis of delayed rejection MCMC methods. Methodology and Computing in Applied Probability 16(4):811–838

[4] Beskos A, Roberts G, Stuart A (2009) Optimal scalings for local Metropolis-Hastings chains on nonproduct targets in high dimensions. Ann Appl Probab 19(3):863–898

[5] Beskos A, Pillai N, Roberts G, Sanz-Serna J, Stuart A, et al (2013) Optimal tuning of the hybrid Monte Carlo algorithm. Bernoulli 19(5A):1501–1534

[6] Girolami M, Calderhead B (2011) Riemann manifold langevin and hamiltonian monte carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73(2):123–214

[7] Haario H, Saksman E, Tamminen J (2001) An adaptive Metropolis algorithm. Bernoulli 7(2):223–242

[8] Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their application. Biometrika 57:97–109

[9] Kim S, Shepard N, Chib S (1998) Stochastic volatility: likelihood inference and comparison with ARCH models. The Review of Economic Studies 65(3):361–393

[10] Mattingly J, Pillai N, Stuart A (2012) Diffusion limits of random walk Metropolis in high dimensions. Annals of Applied Probability 22:881–930

[11] Neal P, Roberts G (2006) Optimal scaling for partially updating MCMC algorithms. Ann Appl Probab 16(2):475–515

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

[13] Roberts GO, Rosenthal JS (1998) Optimal scaling of discrete approximations to Langevin diffusions. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60(1):255–268

[14] Roberts GO, Stramer O (2002) Langevin diffusions and Metropolis-Hastings algorithms. Methodology and computing in applied probability 4(4):337–357

[15] Roberts GO, Gelman A, Gilks WR (1997) Weak convergence and optimal scaling of random walk Metropolis algorithms. Annals of Applied Probability 7:110–120

[16] Sherlock C, Roberts G (2009) Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets. Bernoulli 15(3):774–798

[17] Xifara T, Sherlock C, Livingstone S, Byrne S, Girolami M (2014) Langevin diffusions and the Metropolis-adjusted Langevin algorithm. Statistics & Probability Letters 91:14–19