Title

MALA-within-Gibbs Samplers for High-Dimensional Distributions with Sparse Conditional Structure

Journal

SIAM Journal on Scientific Computing, 42(3)

ISSN

1064-8275

Authors

Tong, XT
Morzfeld, M
Marzouk, YM

Publication Date

2020

DOI

10.1137/19m1284014

Peer reviewed

MALA-within-Gibbs samplers for high-dimensional distributions with sparse conditional structure

X.T. Tong, M. Morzfeld and Y.M. Marzouk

August 27, 2019

Abstract

Markov chain Monte Carlo (MCMC) samplers are numerical methods for drawing samples from a given target probability distribution. We discuss one particular MCMC sampler, the MALA-within-Gibbs sampler, from the theoretical and practical perspectives. We first show that the acceptance ratio and step size of this sampler are independent of the overall problem dimension when (*i*) the target distribution has sparse conditional structure, and (*ii*) this structure is reflected in the partial updating strategy of MALA-within-Gibbs. If, in addition, the target density is block-wise log-concave, then the sampler's convergence rate is independent of dimension. From a practical perspective, we expect that MALA-within-Gibbs is useful for solving high-dimensional Bayesian inference problems where the posterior exhibits sparse conditional structure at least approximately. In this context, a partitioning of the state that correctly reflects the sparse conditional structure must be found, and we illustrate this process in two numerical examples.

1 Introduction

Markov chain Monte Carlo (MCMC) samplers are numerical methods for drawing samples from an arbitrary "target" probability distribution whose density is known up to a normalizing constant. Generically, a Metropolis-Hastings MCMC sampler proposes a move by drawing from a proposal distribution and accepts or rejects the move with a probability that ensures that the stationary distribution of the Markov chain is the target distribution.

To design or chose a sampler for a given distribution, one typically considers the following three criteria. First, the type of proposal distribution is chosen based on how much information about the target distribution is available. For example, the Metropolis adjusted Langevin algorithm (MALA) requires derivatives of the target, while the random walk Metropolis (RWM) algorithm does not. Second, the "step size," which controls how far the proposed sample strays from the current MCMC state, needs to be tuned. Put simply, too large a step size leads to poor mixing because the acceptance probability is too low; too small a step size leads to a large acceptance probability, but the mixing of the chain is poor because a large number of steps are required to produce an effectively independent sample. Step size tuning must find a practical solution to this trade-off, and is problem dependent. Optimal or practical choices of the step size may depend, among other things, on the choice of proposal distribution, the computational resources available, the (apparent or effective) dimension of the problem, and the overall desired accuracy of the MCMC computation. Lastly, in an n-dimensional problem, one can propose an n-dimensional update via an n-dimensional proposal, or one can propose, at each step in the chain, an update for an n/m-dimensional "block" of variables. Such samplers are called "within-Gibbs" samplers,

"partially updating MCMC," "component-wise MCMC," or "partial resampling algorithms"; see, e.g., [4, 23, 28].

The distributions one wishes to sample by MCMC are often high dimensional. Yet the convergence of MCMC samplers often slows in high dimensions, to the extent that calculations become practically infeasible. Our main motivation for this work is that, while it is certainly difficult to sample "generic" high dimensional distributions, distributions that exhibit certain special structure can be feasible to sample, independent of their dimension, provided that the sampler exploits this structure. Examples of samplers in the current literature that leverage various special problem structures are given in Section 4. In this paper, we focus on high-dimensional sampling via the MALA-*within-Gibbs sampler* in the presence of *sparse conditional structure*.

We define sparse conditional structure in Section 3 via the Hessian of the logarithm of the target density. In the special case of a Gaussian target distribution, sparse conditional structure is equivalent to the precision matrix being sparse. More generally, sparse conditional structure is equivalent to the existence of many conditional independence relationships, or the distribution being Markov with respect to a sparse graph [24]. We prove in Section 3 that the partial updating strategy of MALA-within-Gibbs, with carefully defined updates that make use of the sparse conditional structure, leads to acceptance ratios that depend on the dimension of the block-update but are *independent* of the overall dimension. We further show that MALA-within-Gibbs converges with a *rate independent of the dimension* if the target distribution is block-wise log-concave).

We then discuss MALA-within-Gibbs from a practical perspective in Section 5. In this context it is important to realize that the sparse conditional structure may become apparent only after a suitable change of coordinates. For MALA-within-Gibbs to be an effective sampler, we thus need a means of discovering these coordinates, or, equivalently, identifying sparse conditional structure. We expect that many Bayesian inference problems, see, e.g., [2, 14, 32], are naturally formulated in coordinates that exhibit sparse conditional structure, but we also consider an example where a coordinate transformation is required to reveal sparse conditional structure. We further discuss the overall computational efficiency of the MALA-within-Gibbs approach and raise the issue that the partial updating of MALA-within-Gibbs requires m simulations of the numerical model per sample, m being the number of blocks. This implies that there may be an optimal, problem-dependent choice for the dimensionality of the update that can lead to significant computational savings. We explore all of the above issues numerically by applying MALA-within-Gibbs to two well known test problems: a log-Gaussian Cox point process [21,25] and an elliptic PDE inverse problem [3, 19, 27, 30, 39, 40]. We further compare the computational efficiency of MALA-within-Gibbs to the efficiencies of other samplers including MALA, pCN [16], and manifold MALA (MMALA) [21].

2 Notation, assumptions and background

We consider probability distributions with density functions $c\pi(\mathbf{x})$, where c is an unknown normalization constant and π is a known function. We partition the n-dimensional vector $\mathbf{x}$ into m blocks, $\mathbf{x}_1, \ldots, \mathbf{x}_m$, where the subscripts are called "block indices." Note that the blocks $\mathbf{x}_j$ are not necessarily consecutive elements of the vector $\mathbf{x}$.

2.1 Notation

MALA will require gradients of the logarithm of the target density π, which we write as $\mathbf{v}(\mathbf{x}) = \nabla_{\mathbf{x}} \log \pi(\mathbf{x})$. Similarly, we sometimes write derivatives of π with respect to the blocks as $\mathbf{v}_j(\mathbf{x}) = \nabla_{\mathbf{x}_j} \log \pi(\mathbf{x})$. We write $\nabla^2_{\mathbf{x}_i,\mathbf{x}_j}$ to denote second derivatives with respect to blocks i and j, i.e.,

$\nabla^2_{\mathbf{x}_i,\mathbf{x}_j} \log \pi(\mathbf{x})$ is a matrix of size $\dim(\mathbf{x}_i) \times \dim(\mathbf{x}_j)$. Throughout this paper, we use the Euclidean norm for vectors, i.e., $\|\mathbf{x}\| = \sqrt{\mathbf{x}^T\mathbf{x}}$, where superscript T denotes a transpose, and the l_2-operator norm, $\|\mathbf{A}\|$, for matrices. We write $\mathbf{A} \preceq \mathbf{B}$, where $\mathbf{A}$ and $\mathbf{B}$ are two $n \times n$ matrices, when the matrix $\mathbf{A} - \mathbf{B}$ is negative semi-definite. We write $\lambda_{\min}(\mathbf{A})$ for the smallest eigenvalue of the matrix $\mathbf{A}$.

We write conditional densities of one block, $\pi(\mathbf{x}_j|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_m)$, as $\pi(\mathbf{x}_j|\mathbf{x}_{\setminus j})$, i.e, the block index set $\setminus j = \{1, 2, \ldots, j-1, j+1, \ldots, m\}$. More generally, we write $\mathcal{I}$ for a subset of block indices, i.e., $\mathcal{I}$ is a subset of $\{1, 2, \ldots, m\}$. The cardinality of $\mathcal{I}$ will be denoted as $|\mathcal{I}|$. We denote the complement of $\mathcal{I}$ by $\mathcal{I}^c$, i.e., $\mathcal{I}^c$ is the subset of $\{1, 2, \ldots, m\}$ which excludes the block indices in $\mathcal{I}$.

An important concept we will use repeatedly is conditional independence. Conditional independence means that conditioning block i on all but a few other blocks is irrelevant, which we write as

$$\mathbf{x}_j \perp\!\!\!\perp \mathbf{x}_{\mathcal{I}_j^c} \mid \mathbf{x}_{\mathcal{I}_j \setminus \{j\}},$$

where the index set $\mathcal{I}_j$ depends on j and includes the block index j and where $\mathcal{I}_j \setminus \{j\}$ is the index set $\mathcal{I}_j$ with index j removed. In terms of probability distributions, conditional independence means that

$$\pi(\mathbf{x}_j|\mathbf{x}_{\setminus j}) = \pi(\mathbf{x}_j|\mathbf{x}_{\mathcal{I}_j \setminus \{j\}})$$

We assume throughout that $\mathcal{I}_j$ has at most $S \ll m$ elements.

2.2 Assumptions

We assume throughout this paper that $\pi(\mathbf{x})$ has continuous second derivatives and that

(i) the dimension, n, of $\mathbf{x}$ and the number of blocks, m, are large;

(ii) any block $\mathbf{x}_j$ is conditionally independent of most other blocks.

We refer to assumption (ii) as *sparse conditional structure*. This terminology is inspired by linear algebra and Gaussian $\pi(\mathbf{x})$—a Gaussian with sparse conditional structure is characterized by a sparse precision matrix. We make assumption (ii) mathematically more precise in Section 3.1. For simplicity, we assume that n/m (dimension divided by the number of blocks) is an integer.

2.3 Background: MALA and MALA-within-Gibbs

The MALA sampler with n-dimensional updates and step size τ generates a sequence of iterates $\mathbf{x}^k$ by repeating the following two steps, starting from a given $\mathbf{x}^0$:

1. Draw a sample $\tilde{\mathbf{x}}^k$ from the MALA proposal by

$$\tilde{\mathbf{x}}^k = \mathbf{x}^k + \tau \mathbf{v}(\mathbf{x}^k) + \sqrt{2\tau}\boldsymbol{\xi}^k,$$

   where $\boldsymbol{\xi}^k$ is an independent sample from $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$.

2. Accept this proposal with probability

$$\alpha(\mathbf{x}^k, \tilde{\mathbf{x}}^k) = \min\left\{1, \frac{\pi(\tilde{\mathbf{x}}^k)\exp(-\frac{1}{4\tau}\|\mathbf{x}^k - \tilde{\mathbf{x}}^k - \tau\mathbf{v}(\tilde{\mathbf{x}}^k)\|^2)}{\pi(\mathbf{x}^k)\exp(-\frac{1}{4\tau}\|\tilde{\mathbf{x}}^k - \mathbf{x}^k, -\tau\mathbf{v}(\mathbf{x}^k)\|^2)}\right\},$$

   i.e., let $\mathbf{x}^{k+1} = \tilde{\mathbf{x}}^k$ with probability $\alpha(\mathbf{x}^k, \tilde{\mathbf{x}}^k)$, and $\mathbf{x}^{k+1} = \mathbf{x}^k$ with probability $1 - \alpha(\mathbf{x}^k, \tilde{\mathbf{x}}^k)$.

It is straightforward to show that $c\pi(\mathbf{x})$ is the invariant distribution of the Markov chain $\mathbf{x}^k$. Therefore, when $k \to \infty$, $\mathbf{x}^k$ can be viewed as a sample from $c\pi(\mathbf{x})$.

MALA-within-Gibbs is a variation of MALA that uses n/m-dimensional updates. We use superscripts to index "time" in the Markov chain (see above), and subscripts to index the blocks. Thus, starting with a vector $\mathbf{x}^0$ and step size τ, MALA-within-Gibbs iterates the following steps:

1. Set $\mathbf{x}^k = \mathbf{x}^{k-1}$. Repeat steps (a) and (b) below for $j = 1, \ldots, m$ to update all m blocks of $\mathbf{x}^k = [\mathbf{x}_1^k, \ldots, \mathbf{x}_m^k]$.
   (a) Sample a standard Gaussian $\boldsymbol{\xi}_j^k$ of the same dimension as $\mathbf{x}_j$ and use the MALA proposal for the current block $\mathbf{x}_j$:
   $$\tilde{\mathbf{x}}_j^k = \mathbf{x}_j^k + \tau \mathbf{v}_j(\mathbf{x}^k) + \sqrt{2\tau}\boldsymbol{\xi}_j^k, \tag{2.1}$$
   (b) Define $\tilde{\mathbf{x}}^k = [\mathbf{x}_1^k, \ldots, \mathbf{x}_{j-1}^k, \tilde{\mathbf{x}}_j^k, \mathbf{x}_{j+1}^k, \ldots, \mathbf{x}_m^k]$, i.e., $\tilde{\mathbf{x}}^k$ is equal to $\mathbf{x}^k$, except at its j-th block. Compute the block acceptance ratio $\alpha_j(\mathbf{x}^k, \tilde{\mathbf{x}}^k)$. Set $\mathbf{x}^k$ be $\tilde{\mathbf{x}}^k$ with probability $\alpha_j(\mathbf{x}^k, \tilde{\mathbf{x}}^k)$, else $\mathbf{x}^k$ maintains its value.
2. Increase the time index from k to $k+1$ and go to 1.

As before, it is straightforward to verify that the target $c\pi(\mathbf{x})$ is the invariant distribution of the MALA-within-Gibbs iterates. The partial updating can be derived from applying MALA within a Gibbs iteration (hence the name), i.e., with target distributions $\pi(\mathbf{x}_j|\mathbf{x}_{\setminus j})$.

3 Dimension independent acceptance and convergence rate

The MALA and MALA-within-Gibbs samplers can, in principle, be used for arbitrary target distributions, but in generic high-dimensional problems we expect that convergence is slow. In high-dimensional problems with sparse conditional structure, however, MALA-within-Gibbs can be effective if the partitioning of $\mathbf{x}$, defining the partial updates, is chosen in accordance with the sparse conditional structure. With a suitable partial updating strategy, we show that the step size and the acceptance ratio of MALA-within-Gibbs (within each block) can be made independent of the overall dimension. We then show, under additional assumptions of block-wise log-concavity, that MALA-within-Gibbs converges to the target distribution at a dimension-independent rate. The proofs of the propositions and the theorem can be found in the Appendix.

3.1 Dimension independent acceptance under sparse conditional structure

To simplify the proofs, the conditional independence assumption is formulated in terms of the gradient $\mathbf{v}(\mathbf{x})$.

Assumption 3.1 (Sparse conditional structure)**.** *For $\pi(\mathbf{x})$ and the partition $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_m)$, there are constants S and q independent of n, so that*

(i) The dimension of each block $\mathbf{x}_j$ is bounded by q.

(ii) For each block index $j \in 1, \ldots, m$, there is a block index set $\mathcal{I}_j \subset \{1, \ldots, m\}$ with $j \in \mathcal{I}_j$ and cardinality $|\mathcal{I}_j| \leq S$ so that

$$\nabla^2_{\mathbf{x}_k, \mathbf{x}_j} \log \pi(\mathbf{x}) = \nabla_{\mathbf{x}_k} \mathbf{v}_j(\mathbf{x}) = \mathbf{0}, \quad \textit{if} \quad k \notin \mathcal{I}_j.$$

Note that (i) is trivial because we deal with finite dimensional problems and note that (ii), by Lemma 2 of [37], is equivalent to $\mathbf{x}_j \perp\!\!\!\perp \mathbf{x}_{\mathcal{I}_j^c} \mid \mathbf{x}_{\mathcal{I}_j \setminus \{j\}}$ if the density is strictly positive and smooth. In other words, Assumption 3.1 is equivalent to the assumption of sparse conditional structure, as described earlier, but the formulation in terms of gradients is easier to use in our proofs.

Whether or not Assumption 3.1 is satisfied for a given target distribution depends, to a large extent, on how the blocks of $\mathbf{x}$ are defined. Using physical insight into the problem, it is often possible to group components of $\mathbf{x}$ such that Assumption 3.1 is satisfied or approximately satisfied. We discuss this issue more in Section 5 below, but it is important to understand that Assumption 3.1 essentially requires a "good understanding" of the target distribution and that the results we derive under this assumption make use of the fact that one understands and leverages conditional independencies among the components of $\mathbf{x}$.

We also assume that the gradient $\mathbf{v}(\mathbf{x})$ and its derivatives are bounded.

Assumption 3.2 (Bounded vector fields)**.** *The vector field $\mathbf{v}_j(\mathbf{x}) = \nabla_{\mathbf{x}_j} \log \pi(\mathbf{x})$, for $j = 1, \ldots, m$, and its first derivatives are bounded, i.e., there exist constants M_v and H_v, independent of the overall dimension n, such that*

$$\|\mathbf{v}_j(\mathbf{x})\| \leq M_v, \quad \|\nabla_{\mathbf{x}_i} \mathbf{v}_j(\mathbf{x})\| \leq H_v, \quad \|\nabla_{\mathbf{x}_j} \mathbf{v}_j(\mathbf{x}) - \nabla_{\mathbf{z}_j} \mathbf{v}_j(\mathbf{z})\| \leq H_v \|\mathbf{x} - \mathbf{z}\|.$$

By Assumption 3.1, $\mathbf{v}_j(\mathbf{x})$ has no dependence on $\mathbf{x}_{\mathcal{I}_j^c}$, so one can write it as $\mathbf{v}_j(\mathbf{x}_{\mathcal{I}_j})$. If the support of $\pi(\mathbf{x})$ is bounded, then $\mathbf{v}_j(\mathbf{x})$ having no dependence on $\mathbf{x}_{\mathcal{I}_j^c}$, along with the fact that the dimension of each block $\mathbf{x}_i$ is at most q, often yields Assumption 3.2. Unbounded support is more complicated. A Gaussian, for example, violates Assumption 3.2 because the norm of the gradient is not bounded. This boundedness assumption, however, is made for simplicity and may not be required in practice. More sophisticated constructions may be used in the future to relax this assumption and to derive more general results.

Under Assumptions 3.1 and 3.2, the following proposition shows that the step size and the acceptance ratio of MALA-within-Gibbs are independent of the overall dimension.

Proposition 3.3 (Block acceptance)**.** *Suppose $c\pi(\mathbf{x})$ is the density of a distribution with sparse conditional structure (Assumption 3.1) and that, in addition, Assumption 3.2 holds. There is a constant M, independent of the number of blocks m, so that, for any given state $\mathbf{x} \in \mathbb{R}^n$, the block acceptance ratio $\alpha_j(\mathbf{x}^k, \tilde{\mathbf{x}}^k)$ is bounded below by*

$$\mathbb{E}[\alpha_j(\mathbf{x}^k, \tilde{\mathbf{x}}^k)] \geq 1 - M\sqrt{\tau}.$$

for all blocks $j \in \{1, \ldots, m\}$.

Proposition 3.3 follows directly from Lemma A.2, which we prove in Section A. The dimension independent block acceptance ratio is intuitive. Partial updating implies that the proposed updates are, by design, low-dimensional: their dimension depends on the block size, q, but is independent of the number of blocks, m, or the overall dimension $n = m \cdot q$. Thus, only the dimension of the update controls the block acceptance ratio. The overall number of low-dimensional updates, which defines the overall dimension, is irrelevant.

It might seem that Proposition 3.3 contradicts earlier results on optimal scalings of MALA step sizes, where the optimal step size decreases with dimension at a well-understood rate [5, 6, 33, 34]. These earlier scaling results, however, do not assume sparse conditional structure. Thus, in general, the optimal step size of MALA and MALA-within-Gibbs should decrease with dimension, but *if the target distribution has sparse conditional structure and if, in addition, this structure is used*

in the block updates, then the step size (and acceptance ratio) can be independent of the overall dimension.

Assumption 3.1 ensures that the sparse conditional structure of the target is exploited by the MALA-within-Gibbs sampler. We thus assume away any difficulties of discovering sparse conditional structure, but we discuss practical aspects of this assumption in Section 5, including a brief discussion of what happens when the assumptions are only "nearly" met. We also emphasize that we have no claims at "optimal" step sizes of MALA-within-Gibbs—we merely show that the step size need *not* decrease with dimension to ensure a constant average block acceptance ratio. Moreover, "local" tunings as discussed in [4] may further improve efficiency, but we do not pursue such ideas here.

Partial updating of MALA has also been considered in [28], where the conclusion is that the updates in MALA-within-Gibbs should be high-dimensional. Again, this is true in general, but if the target has sparse conditional structure, the dimensionality of the updates may depend on this structure. We revisit this issue in Section 5, where we also bring up a trade-off between the block size and computational requirements that may increase with the number of blocks. One can also perform "random" partial updates, i.e., choosing at random which components of $\mathbf{x}$ are next updated. Asymptotically, MALA-within-Gibbs with a random partial updating strategy converges to the target distribution, but we expect that the convergence will be slow for problems with sparse conditional structure because this structure is *not* used by random partial updates.

3.2 Dimension independent convergence rate

We have shown above that the step size and acceptance ratio of MALA-within-Gibbs can be independent of dimension if the sparse conditional structure of the target distribution is known and used via a suitable partition of the variables during the within-Gibbs moves. This is not enough to guarantee fast convergence of MALA-within-Gibbs. To study the convergence rate of MALA-within-Gibbs, we require, as an additional assumption, that the target distribution be unimodal and block-wise log-concave (see below for a definition). The reason is that difficulties with MCMC that arise from high dimensionality or multi-modality are independent of each other: if the target distribution has multiple modes, a large number of samples may be required even if the dimension is small. We focus on aspects of high-dimensional problems with a single mode.

The additional assumption we need in our proof (see Section A) is "block-wise log-concavity." To define block-wise log-concavity, we first construct an $m \times m$ matrix $\mathbf{H}(\mathbf{x})$, where m is the number of blocks, with the following properties.

Definition 3.4. *A symmetric $m \times m$ matrix function $\mathbf{H}(\mathbf{x})$ with entries $H_{j,i}(x)$ is* uniformly bounded and negative *if there are strictly positive constants H_v and λ_H such that for all j, i and all $\mathbf{x}$,*

$$|H_{j,i}(\mathbf{x})| \leq H_v, \quad \lambda_{\max}(\mathbf{H}(\mathbf{x})) \leq -\lambda_H < 0.$$

As a simple example, a constant symmetric negative definite matrix is uniformly bounded and negative. Block-wise log-concavity can now be formulated as follows.

Assumption 3.5. *A probability density $c\pi(\mathbf{x})$ is block-wise log-concave with block size m if there exists a $m \times m$ uniformly bounded and negative matrix $\mathbf{H}(\mathbf{x})$ such that*

i) $\nabla^2_{\mathbf{x}_j,\mathbf{x}_j} \log \pi(\mathbf{x}) \preceq H_{j,j}(\mathbf{x})\, \mathbf{I}$, *where $\mathbf{I}$ is the identity matrix of size $dim(\mathbf{x}_j) \times dim(\mathbf{x}_j)$;*

ii) the off-diagonal elements bound the conditional dependence between blocks, i.e. $\|\nabla^2_{\mathbf{x}_j,\mathbf{x}_i} \log \pi(\mathbf{x})\| \leq H_{j,i}(\mathbf{x})$ *for all $i \neq j$.*

Setting	q=1	q=2	q=4	q=8	q=16	q=32
$l = 2$	-0.71	-1.28	-1.44	-1.28	-0.71	**0.24**
$l = 1$	**0.04**	-0.21	-0.29	-0.21	**0.04**	**0.46**
$l = 0.5$	**0.62**	**0.54**	**0.52**	**0.54**	**0.62**	**0.76**

Table 1: Block-wise log-concavity with $\mathbf{H}$ as defined in the text for different block sizes q and different correlation length scales l. Positive numbers, highlighted in bold, indicate block-wise log-concavity.

Note that if the dimension of the blocks is $q = 1$, so that $m = n$, and if the Hessian of $\log \pi(\mathbf{x})$ is diagonally dominant, then $\mathbf{H}(\mathbf{x})$ can be taken as the Hessian of π, with all off-diagonal entries replaced by their absolute value. Further note that block-wise log-concavity is a stronger assumption than log-concavity—the function $\pi(\mathbf{x})$ can be log-concave but *not* block-wise log-concave (see example below). On the other hand, a distribution that is block-wise log-concave, for any block size, is also log-concave.

As an illustration, we consider a Gaussian distribution for $\mathbf{x} = [x_1, \ldots, x_{64}]$ with mean zero and covariance matrix $\mathbf{C}$ with elements

$$[C]_{i,j} = \exp\left(-\frac{1}{2l}|i-j|\right), \quad i, j = 1, \ldots, 64.$$

Interpreting this Gaussian as a discretization of a 1D random field with exponential covariance kernel (and discretization $\Delta x = 1$), the quantity l is a correlation length scale. If l is small, only those components of $\mathbf{x}$ that are near each other in the 1D domain are significantly correlated. This suggests partitioning $\mathbf{x}$ based on "neighborhoods" in the 1D domain, which correspond to consecutive elements of $\mathbf{x}$. For example, the block size $q = 4$ results in $m = 16$ blocks

$$\mathbf{x}_1 = [x_1, \ldots, x_4], \quad \mathbf{x}_2 = [x_5, \ldots, x_8], \quad \ldots, \quad \mathbf{x}_{16} = [x_{61}, \ldots, x_{64}].$$

Recall that, for Gaussian distributions, the precision matrix $\mathbf{P}$ is equal to $-2\nabla^2 \log \pi$, which suggests to construct the matrix $\mathbf{H}(x) \in \mathbb{R}^{m \times m}$ by

$$H_{i,i}(\mathbf{x}) \equiv -\lambda_{\min}(\mathbf{P}_{i,i}), \quad H_{i,j}(\mathbf{x}) \equiv \|\mathbf{P}_{i,j}\|.$$

Here, $\mathbf{P}_{i,j}$ is the i,j-th $q \times q$ sub-block of $\mathbf{P}$ with indices corresponding to the blocks $\mathbf{x}_i$ and $\mathbf{x}_j$. For example, with $q = 4$, $\mathbf{P}_{1,2}$ is a sub-block of $\mathbf{P}$ consisting of rows 1-4 and columns 5-8. Assumption 3.5 is then equivalent to assuming that $\mathbf{H}$ is negative definite, i.e., $\lambda_{\min}(-\mathbf{H}) > 0$. We can numerically check this condition by computing eigenvalues of $\mathbf{H}$. Table 1 lists values of $\lambda_{\min}(-\mathbf{H})$ for varying correlation length scales l and block sizes q.

We note that while the Gaussian is log-concave for any l, block-wise log-concavity depends on the length scale l and the "size" of the blocks q. If l is large, only large blocks lead to block-wise log concavity (with $q = 64$ guaranteeing log-concavity *and* block-wise log-concavity). If the correlation is (essentially) confined to "small" neighborhoods, i.e., if l is small, then small block sizes q also lead to block-wise log-concavity.

With the definition of block-wise log-concavity, we can now state a theorem about the dimension-independent convergence rate of MALA-within-Gibbs. The proof is given in Section A.

Theorem 3.6. *Under Assumptions 3.1, 3.2, and 3.5, for any $\delta > 0$, there exists a $\tau_0 > 0$ independent of the number of blocks m, so that when the step size $\tau < \tau_0$, we can couple two MALA-*

within-Gibbs samples $\mathbf{x}^k$ *and* $\mathbf{z}^k$*, such that*

$$\sum_{i=1}^{m} \left(\mathbb{E}\|\mathbf{x}_i^k - \mathbf{z}_i^k\| \right)^2 \leq (1 - (1-\delta)\lambda_H \tau)^{2k} \sum_{i=1}^{m} \left(\mathbb{E}\|\mathbf{x}_i^0 - \mathbf{z}_i^0\| \right)^2 .$$

In particular, one can let $\mathbf{z}^0 \sim \pi$*. It follows that* $\mathbf{z}^k \sim \pi$*, which in turn shows that* $\mathbf{x}^k$ *converges to* π *geometrically fast.*

Theorem 3.6 indicates that MALA-within-Gibbs can be a fast sampler for high-dimensional problems if (i) the target distribution has sparse conditional structure and this structure is used by the MALA-within-Gibbs sampler; and (ii) the target distribution is block-wise log-concave.

Block-wise log-concavity implies that the target has only one mode. Assuming block-wise log-concavity thus allows us to study computational barriers due to high dimensionality without requiring that we simultaneously consider challenges due to multi-modality. Notably, block-wise log-concavity is more restrictive than log-concavity, which also implies that the target is unimodal. We use block-wise log-concavity here to gain stronger control over the coupling between blocks in the analysis. Ultimately, a less restrictive assumption (e.g., log-concavity) may be preferable and one may view our results as a first step towards a full understanding of how MALA-within-Gibbs can operate effectively in high-dimensional problems.

Theorem 3.6 also has connections to previous work on Gibbs samplers for Gaussian distributions with sparse conditional structure [26]. In particular, Theorems 3.1 and 3.2 of [26] show dimension independent convergence of a Gibbs sampler for Gaussian distributions. By interpreting the block-wise log-concavity assumption as a generalization of the Gaussian assumptions in Theorems 3.1 and 3.2 of [26], one can understand Theorem 3.6 as a generalization of this result to a "within-Gibbs" sampler for non-Gaussian distributions.

4 Discussion of efficient samplers in high dimensions

We suggested earlier that sampling *generic* high-dimensional distributions is difficult, but if the target distribution has a special structure, then efficient samplers can be constructed. One example are Gaussian distributions. Gaussians can be sampled efficiently even if their dimension is large, either by direct samplers (using techniques from numerical linear algebra for computing matrix square roots) or by MCMC, using analogies between Gibbs samplers and linear solvers to construct matrix splittings for accelerated sampling; see, e.g., [20, 29].

There are also several routes to making MCMC samplers effective for high-dimensional distributions that are not Gaussian, and we discuss some of them here in relation to our results. One issue with Metropolis-Hastings samplers is that their step size, τ, needs to be tuned. This requires in particular that τ must decrease with dimension, n. Optimal scalings of τ with dimension for various MCMC samplers have been derived: for RWM $\tau_{\text{opt}} = O(n^{-1})$, for MALA $\tau_{\text{opt}} = O(n^{-1/3})$, and for HMC $\tau_{\text{opt}} = O(n^{-1/4})$. Fixing the acceptance ratio with small τ, however, comes with a price: the acceptance ratio may be large, but all accepted steps are small (on average on the order of τ), so that the sampler moves often, but slowly. As a rule of thumb, it takes about $O(1/\tau)$ iterations to move through the support of the target distribution (putting aside issues of reaching stationarity [15]). For very large dimensions, many MCMC samplers are thus slow to converge. These results hold for *general* target distributions. Even if the analyses that lead to the optimal scalings rely on certain assumptions, e.g., that the target measure is of product type, this problem structure is not directly used by the samplers. Our results on dimension independent step size and

convergence rates hold only for target distributions with sparse conditional structure and when this structure is explicitly used by the MALA-within-Gibbs sampler (via Assumption 3.1).

Another strategy for effective sampling of high-dimensional distributions applies if the parameters $\mathbf{x}$ can be decomposed as $\mathbf{x} = (\mathbf{y}, \mathbf{z})$, where $\mathbf{z}$ is low dimensional and, conditioned on $\mathbf{z}$, there are fast (direct) samplers for $\mathbf{y}$; see., e.g., [9, 10, 12]. In Bayesian inverse problems, this structure can often be identified by a suitable choice of basis or reparameterization, as in [18,38,42]: typically $\mathbf{z}$ represents directions where the posterior departs significantly from the prior, while $\mathbf{y}$ represents prior-dominated directions of the parameter space, which can even be (approximately) independent of $\mathbf{z}$. Note that the MALA-within-Gibbs sampler does not require fully, partially, or conditionally Gaussian target distributions, but our analysis of its efficiency requires sparse conditional structure and block-wise log-concavity.

The theory of function-space MCMC also has led to effective MCMC methods for a class of high-dimensional Bayesian inverse problems; see, e.g., [16, 22, 40]. The basic idea is to exploit the fact that an "effective dimension" can be small even if the apparent parameter dimension is large. This happens in particular when high dimensionality comes from the refinement of resolution (e.g., when the parameters represent the discretization of a function), but when the number of observations remains relatively constant. For example, Fourier modes that have vanishing influence on the observations should be easy to sample, no matter how numerous they are. In this case, the performance of MCMC samplers can be made independent of the apparent parameter dimension. There are now many variations of such "discretization invariant" MCMC samplers [7, 13, 18, 35]. with applications discussed in [8,31]. The notion of high dimensionality considered here differs from that assumed in function space MCMC. For Gaussian target distributions, for example, a function-space MCMC proposal that leaves the prior invariant can be efficient if the posterior covariance is a low-rank update of the prior covariance. The assumption of sparse conditional structure, on the other hand, implies for Gaussian target distributions a sparse target precision that is potentially high rank, and allows for high-rank updates from prior to posterior (see [26] for a more thorough discussion).

5 Practical considerations and numerical experiments

Our results on dimension independent step size, acceptance probabilities, and convergence rates hold under precise mathematical assumptions of sparse conditional structure and log-concavity (see Section 3). We now focus on posterior distributions that arise in Bayesian inference problems, because of their practical importance and because we anticipate that the assumption of sparse conditional structure is often satisfied in such problems. We also demonstrate how to use MALA-within-Gibbs in two numerical examples, and discuss and compare the computational costs of MALA-within-Gibbs and other MCMC samplers.

5.1 Posterior distributions with sparse conditional structure

The Bayesian problem setup is as follows. Let $\mathbf{x}$ be an n-dimensional vector endowed with a prior probability density $\pi_0(\mathbf{x})$. In many problems, $\mathbf{x}$ arises from a discretization of a spatially distributed quantity (i.e., a "field") and, for that reason, is high dimensional. The prior reflects assumptions about the smoothness of the field and is often assumed to be Gaussian with a known mean and covariance. A computational model, $\mathcal{M}(\mathbf{x})$, maps $\mathbf{x}$ to observations $\mathbf{y}$. Typically, the model is nonlinear and the number of observations is less than the dimension of $\mathbf{x}$. Any model errors are

represented by a random variable $\boldsymbol{\varepsilon}$ and, often, model errors are additive, i.e.,

$$\mathbf{y} = \mathcal{M}(\mathbf{x}) + \boldsymbol{\varepsilon}. \tag{5.1}$$

The distribution of $\boldsymbol{\varepsilon}$ is assumed to be known (often Gaussian with mean zero and diagonal covariance matrix). Equation (5.1) defines a likelihood $\pi_l(\mathbf{y}|\mathbf{x})$ and the likelihood and prior jointly define the posterior distribution

$$\pi(\mathbf{x}|\mathbf{y}) \propto \pi_0(\mathbf{x})\pi_l(\mathbf{y}|\mathbf{x}).$$

Sparse conditional structure arises naturally in Bayesian posterior distributions when (*i*) the parameters $\mathbf{x}$ are high dimensional, but not all components of $\mathbf{x}$ have significant statistical interactions; and (*ii*) each observation is informative for only a small subset of the components of $\mathbf{x}$ (see also [26]). Put differently, we assume that the prior has sparse conditional structure and that the observations do not significantly densify the conditional structure of the prior. This happens in many geophysical applications, e.g., in numerical weather prediction (NWP), where the posterior distribution is defined jointly by a global atmospheric model (with dimension $O(10^8)$) and observations of the atmospheric state (typically $O(10^7)$ observations). In a global atmospheric model, each model component stores information about the atmospheric state at a specific location at a given time and each component has significant statistical interactions with nearby components, but not with components that are far away. A discussion of the mathematical mechanisms that lead to this property can be found in [11].

5.2 Implementation of MALA-within-Gibbs

The partitioning of $\mathbf{x}$ into blocks is important for effective sampling of the posterior by MALA-within-Gibbs, because only a "suitable" partition will indeed put the sparse conditional structure to use. We do not have a general strategy to find a suitable partition, but we expect that a workable partition is often intuitive. For example, if $\mathbf{x}$ is defined over a spatial domain (1D, 2D, or 3D) and if correlations are limited to small neighborhoods, then the partitioning should be based on these neighborhoods and the block size should take the correlation lengths scales into account. We demonstrate this process in a numerical example in Section 5.3. Our second example in Section 5.4 demonstrates how to choose a partition for the partial updating based on prior covariances in the absence of a spatial scale.

The partial updating of MALA-within-Gibbs requires m likelihood evaluations per sample. In a typical Bayesian inverse problem, each likelihood evaluation will require a full forward solve with the numerical model $\mathcal{M}$, even if only one block of the model's components is updated. Using the common "effective sample size"

$$N_{\text{eff}} = N_e/\text{IACT}, \tag{5.2}$$

where N_e is the number of MCMC samples (length of the Markov chain) and IACT is the integrated auto correlation time, see, e.g., [36, 41], we can estimate the cost per effective sample by

$$\text{cost per effective sample} = \text{IACT} \times (\#\text{ of blocks}) \times \text{cost of likelihood evaluation}. \tag{5.3}$$

It is now clear that the computational cost of MALA-within-Gibbs grows with the number of blocks, even if IACT is independent of dimension. There is, thus, a "hidden" dependence on dimension since a larger dimensional problem also requires a larger number of blocks (keeping other parameters that define the model unchanged, see examples below). This also means that there is a trade-off for sampling in high-dimensions that may not be easy to resolve: to keep the efficiency high (small IACT), one may want to use a large number of small blocks (with a lower bound on the block size depending on the correlation structure), but on the other hand, one may want to use small number of large blocks to keep the number of model evaluations per sample small.

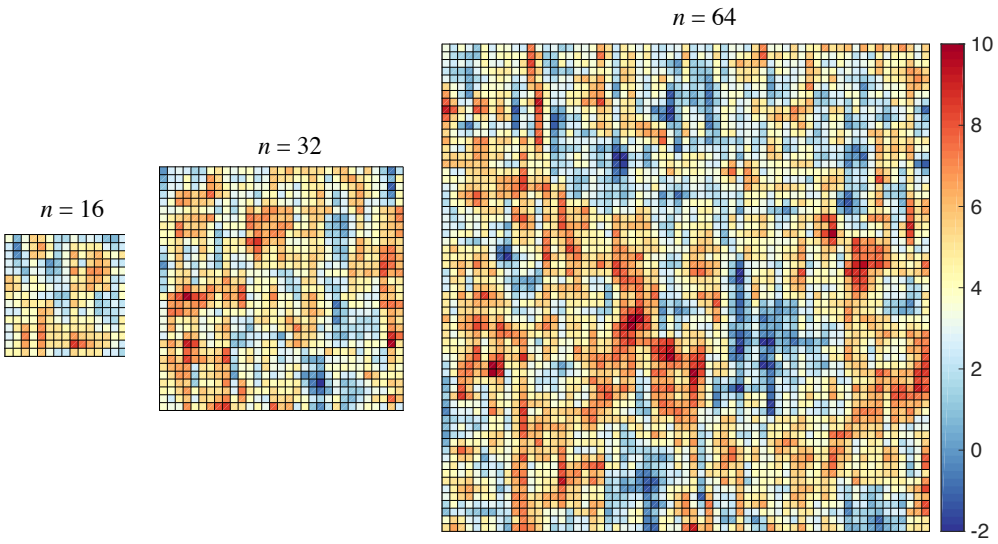

Figure 5.1: True values of X for three problems with increasing domain size L (drawn to scale).

5.3 Numerical illustration 1: Log-Gaussian Cox point processes

We consider inference in a log-Gaussian Cox point process similar to the numerical experiments in [21]. A uniform $N_u \times N_v$ grid, with spacing $\Delta u = \Delta v = 1$, covers the 2D (spatial) domain, $[1, L] \times [1, L]$. The parameter to be inferred is defined over the domain, $\{X_{i,j}, i, j = 1, \ldots, L/\Delta u\}$. Its prior is $\mathcal{N}(\mu\mathbf{1}, \mathbf{B})$, where $\mathbf{B}$ is a discretization of the exponential covariance kernel, i.e.,

$$\mathrm{cov}(X_{s_1,t_1}, X_{s_2,t_2}) = \sigma_s^2 \sigma_t^2 \exp\left(-\frac{1}{2}\frac{|s_1 - s_2|}{l_s} - \frac{1}{2}\frac{|t_1 - t_2|}{l_t}\right),$$

where $\sigma_s^2 = \sigma_t^2 = 2$, $\mu = 4$, $l_s = 2$, $l_t = 4$.

Observations are made at each grid point, denoted by $Y_{i,j}$. The observations are conditionally independent and Poisson distributed with means $\exp(X_{i,j})$. Our goal is to estimate $X_{i,j}$ from $Y_{i,j}$. The prior and likelihood define the posterior distribution

$$\pi(\mathbf{x}|\mathbf{y}) \propto \exp\left(-\frac{1}{2}||\mathbf{B}^{-1/2}(\mathbf{x} - \mu\mathbf{1})||^{\mathbf{2}}\right) \prod_{i,j} \exp\left(Y_{i,j}X_{i,j} - \exp(X_{i,j})\right),$$

where $\mathbf{x}$ is the column stack of $X_{i,j}$, i.e., an $n = L^2$ dimensional vector.

5.3.1 Problem setup

We consider three problems with increasing domain size $L = 16$, $L = 32$, and $L = 64$, leading to sampling problems of dimensions 256, 1024, and 4096. The true values of $X_{i,j}$ for the three domains are shown in Figure 5.1. Note that the (apparent) dimension of the problem ($n = L^2$) and the number of observations (L^2) increase with increasing domain size, but the prior length scales are fixed and short compared to all three domain sizes. Moreover, each observation $Y_{i,j}$ carries information about only one grid point, $X_{i,j}$.

We note that our problem setup is slightly different from that considered in [21], where the means of the Poisson distributions are $\exp([X]_{i,j})/L^2$ (in our notation) and where a different covariance

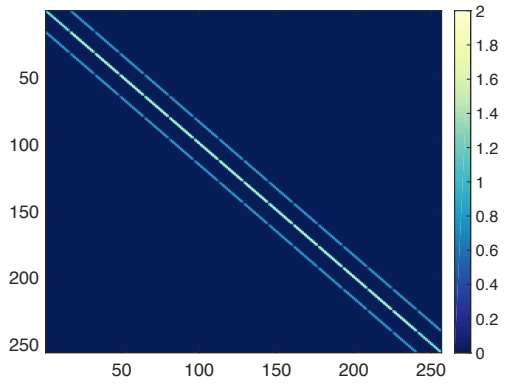

Figure 5.2: Prior precision matrix of the 16×16 problem.

kernel is used to define the prior. The latter is minor. We do not scale the mean values with domain size, L, because we want to study MALA-within-Gibbs on problems with increasing dimension while leaving all other parameters that define the problem unchanged.

The prior precision matrix is sparse, as illustrated in Figure 5.2 for the problem of size 16×16. The prior precision matrices of the larger problems (32×32 and 64×64) have similar sparsity patterns. We now investigate whether Assumptions 3.1 (sparse conditional structure) and 3.5 (block-wise log-concavity) are satisfied in this problem. Sparse conditional structure can verified by inspection: the chosen Gaussian prior is a Markov random field where each pixel has only four neighbors, and the likelihood is purely local, introducing no new dependencies. We can verify this structure more carefully as follows, partitioning the state $\mathbf{x}$ based on 2D-neighborhoods. Recall that the log posterior density is

$$\log \pi(\mathbf{x}|\mathbf{y}) = C - \frac{1}{2}||\mathbf{B}^{-1/2}(\mathbf{x} - \mu\mathbf{1})||^2 + \sum_{i,j} \left(Y_{i,j}X_{i,j} - \exp(X_{i,j})\right), \tag{5.4}$$

where C is a constant whose value is irrelevant. Fixing the block size at $q = n/m$, we find that

$$\nabla_{\mathbf{x}_i,\mathbf{x}_j} \log \pi(\mathbf{x}|\mathbf{y}) = -[\mathbf{B}^{-1}]_{\mathbf{x}_i,\mathbf{x}_j} - \mathbf{1}_{i=j}\mathbf{D}_i, \tag{5.5}$$

where the $q \times q$ matrices $[\mathbf{B}^{-1}]_{\mathbf{x}_i,\mathbf{x}_j}$ are constructed from $\mathbf{B}^{-1}$ based on the blocks $\mathbf{x}_i, \mathbf{x}_j$, and $\mathbf{D}_i$ is a diagonal $q \times q$ matrix with entries being $\exp(X_{k,l})$ for each $X_{k,l}$ in $\mathbf{x}_i$. Due to the sparse structure of the prior precision $\mathbf{B}^{-1}$ (see Figure 5.2), $[\mathbf{B}^{-1}]_{\mathbf{x}_i,\mathbf{x}_j}$ is zero if the blocks i and j are far from each other in the 2D domain. In this case, $\mathbf{1}_{i=j}\mathbf{D}_i = 0$, so that by (5.5), $\nabla^2_{\mathbf{x}_i,\mathbf{x}_j} \log \pi(\mathbf{x}|\mathbf{y})$ is also zero. The problem is thus indeed characterized by sparse conditional structure.

The block-wise log-concavity Assumption 3.5 may not be satisfied in this example. By (5.5), the Hessian is bounded above by $-\mathbf{B}^{-1}$, which suggests to use

$$H_{i,i}(x) \equiv -\lambda_{\min}([\mathbf{B}^{-1}]_{\mathbf{x}_i,\mathbf{x}_i}), \quad H_{i,j}(x) \equiv \|[\mathbf{B}^{-1}]_{\mathbf{x}_i,\mathbf{x}_j}\|, \quad i,j = 1,\ldots,m$$

where $[\mathbf{B}^{-1}]_{\mathbf{x}_i,\mathbf{x}_i}$ and $[\mathbf{B}^{-1}]_{\mathbf{x}_i,\mathbf{x}_j}$ are constructed from $\mathbf{B}^{-1}$, based on the blocks with indices i and j. With this choice, Assumption 3.5 requires that

$$c := \lambda_{\min}(-\mathbf{H}) > 0. \tag{5.6}$$

With this choice of $\mathbf{H}$ and with the length scales $l_s = 2$, $l_t = 4$, the condition in (5.6) is not satisfied, suggesting that the problem is not block-wise log-concave.

5.3.2 MCMC samplers

We apply pCN, simplified manifold MALA (MMALA) [21], and MMALA-within-Gibbs to draw samples from the posterior distributions of the 16×16, 32×32 and 64×64 problems. The MMALA proposal is

$$\tilde{\mathbf{x}}^k = \mathbf{x}^k + \tau \mathbf{M} \nabla \log p(\mathbf{x}^k|\mathbf{y}) + \sqrt{2\tau}\, \mathbf{M}^{1/2} \xi^{k+1}, \quad \xi^{k+1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

where the choice $\mathbf{M} = \mathbf{\Lambda} + \mathbf{B}^{-1}$ turns MALA into simplified manifold MALA. The matrix $\mathbf{\Lambda}$ is diagonal and the ith diagonal element is $[\mathbf{\Lambda}]_{i,i} = \exp(\mu + [\mathbf{B}]_{i,i})$; see [21]. We implement MMALA-within-Gibbs using blocks of size $q = d \times d$ and consider $d = 8, 16, 32, 64$. We emphasize that MMALA-within-Gibbs with a single block, covering the entire domain, is equivalent to the MMALA sampler. For example, if $L = 64$ and $d = 64$, the sampler does not use partial updating and we recover the "usual" MMALA; with $L = 64$ and $d = 16$, we divide the domain into 16 blocks, each of size 16×16. The blocks define a neighborhood of components $X_{i,j}$ of size $d \times d$ on the 2D-domain and are ordered left-to-right and top-to-bottom.

All samplers are initialized at the maximum a posteriori point (MAP) which we find by solving the optimization problem

$$\min_{\mathbf{x}} -\log \pi(\mathbf{x}|\mathbf{y}),$$

using a Gauss–Newton method. We consider various step sizes τ and, for each one, we run pCN to generate 10^5 samples and MMALA or MMALA-within-Gibbs to generate 10^4 samples. We then compute the integrated auto correlation time (IACT) of each pixel using the techniques described in [41]. Note that we use *all* samples (no burn-in) to compute the average acceptance ratios and IACT. We inspected some of the chains and could not identify an apparent transient phase, likely because our initialization point makes the transients negligible.

5.3.3 MCMC results

Results of a MMALA-within-Gibbs sampler with $d = 8$ and step size $\tau = 0.5$ are shown in Figure 5.3. The panels in the top row show the posterior mean (average of all MCMC samples) and the observations $Y_{i,j}$ (on a log-scale). The panels in the bottom row show the posterior variance at each grid point and the observations (on a log-scale) corresponding to the posterior mean. We note a good agreement between the posterior mean and the "true" field (see Figure 5.1), as well as a good agreement between the observations and the reconstructed obervations.

Our tuning of the step-size is illustrated in Figure 5.4, where the average acceptance ratio of MMALA and MMALA-within-Gibbs is plotted as a function of the step sizes we tried for the problem with $L = 64$. The results are qualitatively similar for the problems with $L = 16$ and $L = 32$. We see that the acceptance ratio decreases with step size, but for any fixed step size τ, the acceptance ratios of the within-Gibbs samplers increase when the block sizes are decreased. The reason is that the partial updating of the MMALA-within-Gibbs sampler results in large acceptance ratios for large step sizes independently of the dimension of the problem. Figure 5.4 also shows IACT as a function of the step size. We note that, for a fixed step size, IACT increases with the block size and that the step size that minimizes IACT increases as the block size decreases. Again, the reason is that the partial updating strategy of MMALA-within-Gibbs allows larger steps for smaller blocks, which decreases IACT and accelerates the mixing of the Markov chain.

IACT, averaged over all grid points, and the average acceptance probabilities (averages taken over the MCMC moves) are listed in Table 2. Here, we list results where we fixed the step size for each considered block size to make the resulting average acceptance probabilities comparable. An even better agreement between the acceptance probabilities at each block would require a more

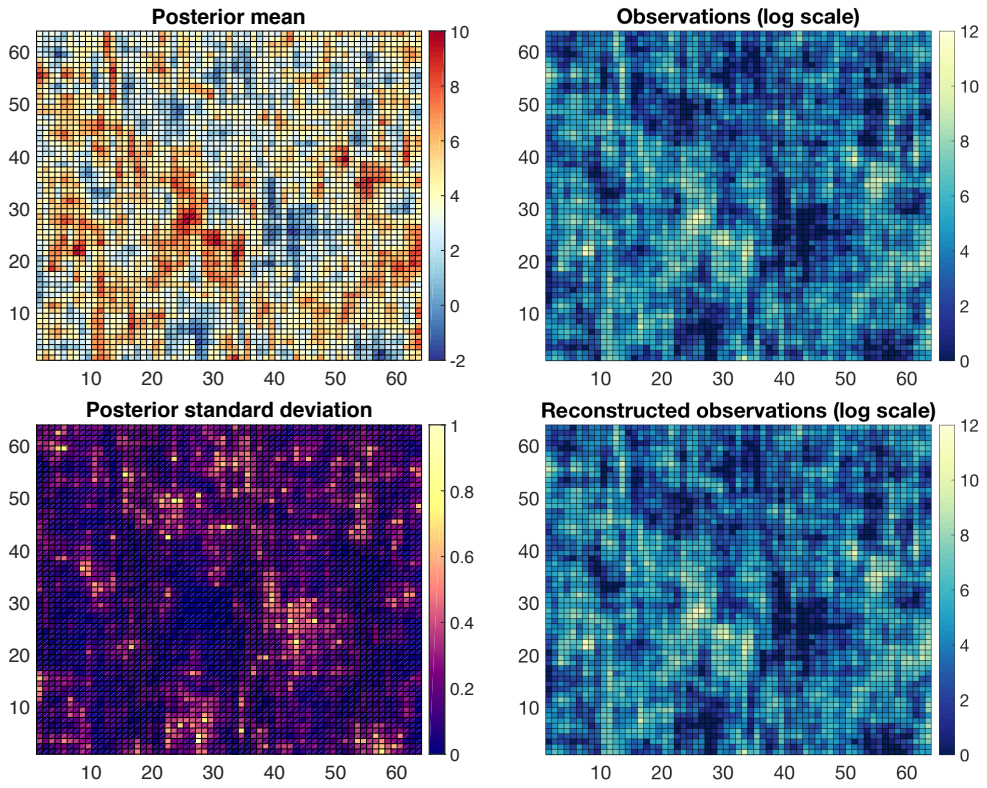

Figure 5.3: Illustration of results obtained by 10^4 samples of a MMALA-within-Gibbs sampler with $d = 8$, and step size $\tau = 0.5$. Top row: posterior mean (left) and observations $Y_{i,j}$ (right). Bottom row: posterior variance at each grid point (left), observations corresponding to posterior mean (right).

careful tuning of the step size, but the step size tuning we carried out is sufficient to make our points and to illustrate the relevant characteristics of the samplers. We note that the IACT of pCN and MMALA with n-dimensional updates increases with L (dimension), and that the IACT of MMALA is lower than that of pCN. MMALA-within-Gibbs yields the same IACT *independently* of the domain size (dimension). For example, with blocks of size $d = 16$, IACT of the 4096-dimensional problem is similar to the IACT of the 256- or 1024-dimensional problems. Moreover, the step size and corresponding acceptance ratios seem to be independent of the overall problem dimension. The numerical experiments thus corroborate our theoretical results on dimension independent convergence of MALA-within-Gibbs, even when the assumption of block-wise log-concavity, required for our proofs, is not satisfied with our choices of block size for MMALA-within-Gibbs.

The dimension-independent convergence, however, does not necessarily imply that MMALA-within-Gibbs is the most efficient sampler for this problem. Using the cost-per-effective-sample in Equation (5.3), it is evident that MMALA is more efficient than MMALA-within-Gibbs (at the block sizes we consider). The cost estimate (5.3), however, assumes that a "full" likelihood evaluation is required for each proposed sample of MMALA-within-Gibbs, which is a conservative estimate. One can easily envision making use of the problem structure during likelihood evaluations in each block. For example, evaluation of the prior term in (5.4) in each block does not require computing the full matrix-vector product $\mathbf{B}^{-1/2}(\mathbf{x} - \mu\mathbf{1})$. One can speed up the computations by only updating the relevant components that are modified in the current block. We did not pursue

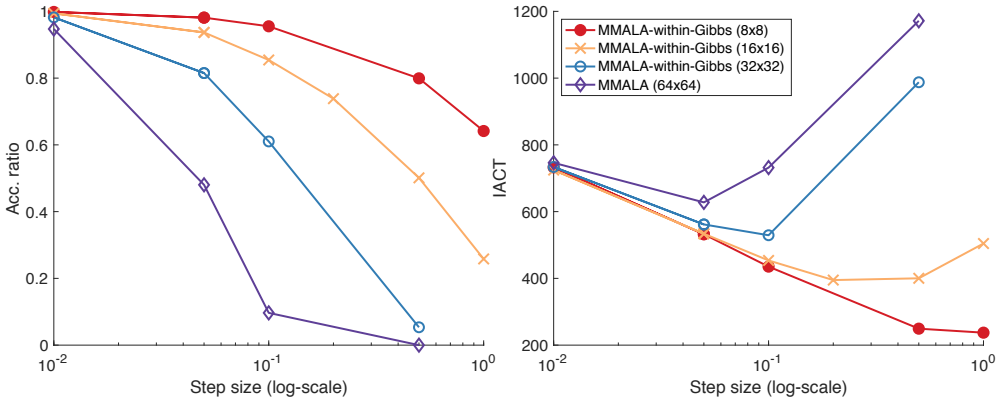

Figure 5.4: Left: average acceptance ratio of MMALA and MMALA-within-Gibbs as a function of the step size for the 64×64 problem. Right: IACT of MMALA and MMALA-within-Gibbs as a function of the step size for the 64×64 problem.

		MMALA-within-Gibbs ($N_e = 10^4$)			
L	pCN ($N_e = 10^5$)	64×64 blocks	32×32 blocks	16×16 blocks	8×8 blocks
16	4626/0.18/0.002	-	-	342/0.95/0.2	204/0.93/0.5
32	5363/0.26/0.002	-	437/0.75/0.1	330/0.73/0.2	203/0.78/0.5
64	6884/0.21/0.001	627/0.48/0.05	529/0.61/0.1	394/0.75/0.2	249/0.80/0.5

Table 2: IACT/average acceptance probability/τ of pCN, MMALA, and MMALA-within-Gibbs for three problems with increasing domain size L (and thus increasing dimension). Note that MMALA-within-Gibbs with block size equal to the domain size corresponds to MMALA.

such ideas because this problem is relatively simple and because our main goal is to demonstrate that MMALA-within-Gibbs can exhibit dimension independence.

5.4 Numerical illustration 2: inverse problems with an elliptic PDE

We consider the PDE

$$-\nabla \cdot (\kappa \nabla u) = g,$$

on a square domain $(s,t) \in [0,1]^2$ with Dirichlet boundary conditions; here u represents a "pressure" field and g is a given source term, which consists of four delta functions (sources) at four locations in the domain. Details on the boundary conditions and source term are given in [27]. The quantity $\kappa > 0$ represents the "permeability" of the medium; we use a log-normal prior for the permeability to enforce the non-negativity constraint. Thus, $K = \log \kappa$ is a Gaussian random field. We set its mean to be zero and employ the covariance kernel

$$k(s_1, t_1; s_2, t_2) = \exp\left(-\frac{(s_1 - s_2)^2}{2l_s^2} - \frac{(t_1 - t_2)^2}{2l_t^2}\right)$$

where (s_1, t_1) and (s_2, t_2) are two points in the square domain and l_s and l_t are correlation length scales. Our goal is to estimate the permeability given 128 noisy observations of the pressure u in the center of the domain. This problem setup is also described in [27]. The inverse problems we consider here differ from those in [27] only in the correlation lengths of the prior, which do not

affect the numerics of the PDE solve, the gradient computations, or the observation and forcing network. We thus refer to [27] for the details of the numerical solution of the PDE, and in particular to Figure 2 of [27] for descriptions of the locations of the forcing terms.

5.4.1 Discretization and problem setups

For computations, we discretize the PDE using a standard finite element method with a uniform grid of 16×16 points (see [27] for details of the discretization). The discretization leads to the algebraic equation

$$\mathbf{A}(\hat{\boldsymbol{\kappa}})\hat{\mathbf{u}} = \hat{\mathbf{g}}, \tag{5.7}$$

where the hat over variables denotes discretized quantities, i.e., $\hat{\boldsymbol{\kappa}}$, $\hat{\mathbf{u}}$, and $\hat{\mathbf{g}}$ are vectors of size $N_u = 256$ and $\mathbf{A}$ is a 256×256 matrix that depends on the permeability $\hat{\boldsymbol{\kappa}}$. We will be computing with the discretized PDE from now on and, for that reason, we drop the hats above all variables. The pressure observations are modeled by the equation

$$\mathbf{y} = \mathbf{H}\mathbf{u} + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}), \tag{5.8}$$

where $\mathbf{H}$ is a $N_y \times N_u$ matrix that has exactly one 1 in each row and picks out every other component of $\mathbf{u}$. The observation noise covariance is set to be $\mathbf{R} = 0.1^2\,\mathbf{I}$.

After discretization, the log-permeability $\mathbf{k}$ is finite dimensional and its prior distribution is the finite dimensional Gaussian $\mathcal{N}(0, \mathbf{B})$. Due to the squared exponential covariance model, $\mathbf{B}$ can be well approximated by a low-rank matrix, i.e.,

$$\mathbf{B} \approx \mathbf{U}_\theta \mathbf{L}_\theta \mathbf{U}_\theta^T, \tag{5.9}$$

where $\mathbf{L}_\theta$ is a $N_\theta \times N_\theta$ diagonal matrix whose diagonal elements are the $N_\theta < N_u$ largest eigenvalues of $\mathbf{B}$ (see [27] for details).

The Gaussian prior for the log-permeability and the likelihood in (5.8) define the posterior distribution $\pi(\mathbf{k}|\mathbf{y}) \propto \pi_0(\mathbf{k})\pi_l(\mathbf{y}|\mathbf{k})$ for the log-permeability:

$$\pi(\mathbf{k}|\mathbf{y}) \propto \exp\left(-\frac{1}{2}\|\mathbf{B}^{-1/2}\mathbf{k}\|^2 - \frac{1}{2}\|\mathbf{R}^{-1/2}(\mathcal{M}(\mathbf{k}) - \mathbf{y})\|^2\right);$$

here $\mathcal{M}$ maps the log permeability to the pressure at observation locations, i.e., $\mathcal{M}(\mathbf{k}) = \mathbf{H}\mathbf{u}(\exp(\mathbf{k}))$, with the $\mathbf{u}$ being the solution to the discretized PDE (5.7).

Since symmetric positive semi-definite matrices can always be diagonalized by a coordinate transformation, we consider the change of variables

$$\boldsymbol{\theta} = \mathbf{L}_{\boldsymbol{\theta}}^{-1/2}\mathbf{U}_{\boldsymbol{\theta}}^T\mathbf{k} \approx \mathbf{B}^{-1/2}\mathbf{k}, \tag{5.10}$$

which leads to the posterior distribution

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \exp\left(-\frac{1}{2}\|\boldsymbol{\theta}\|^2 - \frac{1}{2}\|\mathbf{R}^{-1/2}(\mathcal{M}(\mathbf{k}(\boldsymbol{\theta})) - \mathbf{y})\|^2\right). \tag{5.11}$$

Below, we use MCMC samplers to draw samples from the posterior distribution of $\boldsymbol{\theta}$. The corresponding (log-)permeabilities are computed from posterior samples of $\boldsymbol{\theta}$ via the inverse of the transformation (5.10).

We consider two problem setups, which differ in the correlation lengths of the log-normal prior. The correlation lengths define the dimension of $\boldsymbol{\theta}$ in that the latter is chosen to retain 95% of

	l_s	l_t	$N_{\boldsymbol{\theta}}$
Setup 1	0.4	0.8	30
Setup 2	0.2	0.1	136

Table 3: Correlation lengths and reduced dimensions for Setups 1 and 2.

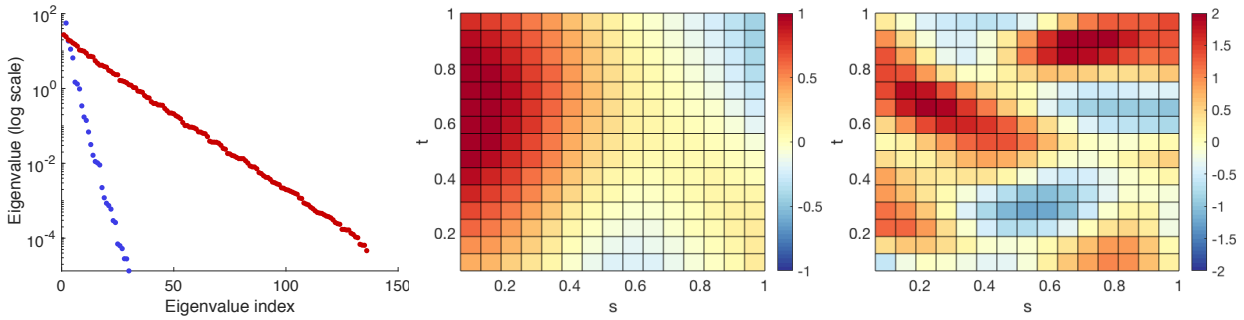

Figure 5.5: Left: eigenvalues of the prior covariance matrix for Setup 1 (blue) and Setup 2 (red). Center: true log-permeability of Setup 1. Right: true log-permeability of Setup 2.

the integrated prior variance. Specifically, if the correlation lengths are short compared to the $[0,1] \times [0,1]$ domain, then the dimension is of $\boldsymbol{\theta}$ is large; if the correlation lengths are large, the dimension of $\boldsymbol{\theta}$ is small. The correlation length scales and implied dimensions of $\boldsymbol{\theta}$ of Setups 1 and 2 are summarized in Table 3.

We illustrate the decay of the prior covariance eigenvalues and the "true" log-permeabilities of Setups 1 and 2 in Figure 5.5. Specifically, we note that the eigenvalues decay more quickly for Setup 1 than for Setup 2 because Setup 1 is characterized by larger correlation length scales than Setup 2. The "true" log-permeabilities of Setups 1 and 2 are random draws from the prior and are shown in the right panels of Figure 5.5. There is more small-scale structure in the log-permeability of Setup 2 than in Setup 1, again due to the shorter prior correlation length scales.

Setup 2 is intended to have a higher dimension than Setup 1, not just in the apparent dimension of $\boldsymbol{\theta}$ but also in the sense of the prior-to-posterior update and hence the influence of the data (i.e., the "effective dimension," as defined in [1]). We achieve this by keeping the domain size fixed while decreasing the correlation lengths, which effectively increases the number of degrees of freedom in the unknown. In the previous log-Cox example, we imposed a similar growth by keeping the correlation lengths fixed but increasing the domain size. Note that, however, in the previous example we also increased the number of observations with the dimension (size of the domain), while in this example, we keep the number of observations fixed. This is a minor issue because in Setup 1, due to the large prior correlation lengths, many of the observations are strongly dependent (i.e., in the prior predictive $\pi(\mathbf{y})$). When the correlation lengths decrease in Setup 2, the number of effectively independent observations increases and thus the relative influence of the likelihood, and hence the effective dimension, increase as well.

The theory we created for the dimension independent convergence of the MALA-within-Gibbs sampler relies on assumptions of sparse conditional structure and block-wise log-concavity. With our choice of prior, the problem does not have sparse conditional structure in (s,t)-coordinates. Yet the coordinate transformation (5.10) produces a sparse conditional structure—indeed complete independence—in the *prior* for the $\boldsymbol{\theta}$-coordinates, which correspond to discretized Karhunen-Loève

(KL) modes. Conditioning on the observations, however, can introduce dependence among the smoother KL modes because an observation at a given (s,t)-location is in principle influenced by all of the modes—due to the nature of the elliptic operator and the KL modes' global support. Conversely: changes in one KL mode can affect the solution everywhere in (s,t)-coordinates. Nonetheless, our experiments, along with various other experiments with this problem found in the literature, suggest that this dependence is weak and that the problem thus has an *approximate* sparse conditional structure in the KL modes. The assumption of block-wise log-concavity is difficult to verify in this example, in either the $\boldsymbol{\theta}$ or (s,t)-coordinates. The reason is that the discretization of the PDE, e.g., in (5.11), makes computations difficult because we do not have second-order adjoints to compute the required Hessian.

5.4.2 MCMC samplers

We use pCN, MALA, and MALA-within-Gibbs to draw samples from the posterior distribution (5.11). Again, we emphasize that MALA-within-Gibbs with a sufficiently large block size is the same as the "usual" MALA without partial updating. The pCN proposal is

$$\tilde{\boldsymbol{\theta}}^{k+1} = \sqrt{1-\beta^2}\boldsymbol{\theta}^k + \beta\boldsymbol{\xi}^{k+1},$$

where $\boldsymbol{\theta}^k$ is the current state of the MCMC and where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{N_{\boldsymbol{\theta}}})$, $\mathbf{I}_{N_{\boldsymbol{\theta}}}$ being the identity matrix of order $N_{\boldsymbol{\theta}}$. The proposed $\tilde{\boldsymbol{\theta}}^{k+1}$ is accepted with probability

$$\alpha_{\text{pCN}} = \min\left(1, \exp\left(\frac{1}{2}\|\mathbf{R}^{-1/2}(\mathcal{M}(\mathbf{k}(\boldsymbol{\theta}^k)) - \mathbf{y})\|^2 - \frac{1}{2}\|\mathbf{R}^{-1/2}(\mathcal{M}(\mathbf{k}(\tilde{\boldsymbol{\theta}}^{k+1})) - \mathbf{y})\|^2\right)\right).$$

We initialize the pCN chain at the MAP, which we find by quasi-Newton optimization (Matlab's fminunc) of the cost function

$$F(\boldsymbol{\theta}) = \log(\pi(\boldsymbol{\theta}|\mathbf{y})) = -\frac{1}{2}\|\boldsymbol{\theta}\|^2 - \frac{1}{2}\|\mathbf{R}^{-1/2}(\mathcal{M}(\mathbf{k}(\boldsymbol{\theta})) - \mathbf{y})\|^2 + C, \tag{5.12}$$

where C is a constant that is irrelevant. We tune the parameter β to obtain minimal IACT. As above, IACT is computed using the techniques and definitions of [41].

The MALA proposal for this problem is

$$\tilde{\boldsymbol{\theta}}^{k+1} = \boldsymbol{\theta}^k - \tau\mathbf{J}^{-1}\nabla_{\boldsymbol{\theta}}F(\boldsymbol{\theta}^k) + \sqrt{2\tau}\mathbf{J}^{-1/2}\boldsymbol{\xi}^{k+1},$$

where $F(\boldsymbol{\theta})$ is as in (5.12) and where $\mathbf{J}$ is the Hessian of F at the MAP. As with pCN, we initialize MALA at the MAP and tune the step size τ of MALA to find a minimal IACT. As in the previous example, we use all samples for our computations (no burn-in).

MALA-within-Gibbs requires that we partition $\boldsymbol{\theta}$ into blocks. Above, we argued that this problem has an approximate sparse conditional structure in the $\boldsymbol{\theta}$ coordinates. For this reason, we use partitions of $\boldsymbol{\theta}$ that group consecutive elements of $\boldsymbol{\theta}$ together. Below, we consider several block sizes and for each one, we initialize MALA-within-Gibbs at the MAP (as before) and tune the step size to achieve a minimal IACT.

5.4.3 MCMC results

Typical results one can obtain via MCMC are shown in Figure 5.6, where we plot an approximation of the posterior mean of the log-permeability and the approximate posterior standard deviations

	Method	Length of chain	IACT	Acc. ratio	Step
Setup 1	MALA-within-Gibbs, $q = 1$	10^4	25	0.43	0.5
	MALA-within-Gibbs, $q = 15$	10^4	141	0.22	0.05
	MALA/MALA-within-Gibbs, $q = 30$	10^5	246	0.44	0.01
	pCN	10^6	6,102	0.24	0.01
Setup 2	MALA-within-Gibbs, $q = 1$	10^3	20	0.38	0.500
	MALA-within-Gibbs, $q = 68$	10^4	367	0.23	0.010
	MALA/MALA-within-Gibbs, $q = 136$	10^5	923	0.23	0.005
	pCN	10^6	28,015	0.45	0.010

Table 4: Summary of simulation results of Setups 1 and 2.

(on the grid) computed via MALA-within-Gibbs. We obtain an approximate posterior mean of the log-permeability, $\mathbf{k}$, from the posterior mean of $\boldsymbol{\theta}$, by mapping $\boldsymbol{\theta}$ to $\mathbf{k}$ via the inverse of (5.10). The approximate posterior mean of $\mathbf{k}$ should be compared to the "true" log-permeability in Figure 5.5.

A summary of the numerical experiments we performed is provided in Table 4. The table lists IACT, step sizes, and average acceptance ratios for the various MCMC samplers. The numbers shown are "tuned," in the sense that we only show results for the step size that leads to minimal IACT (over all step sizes we tried). We note that IACT of pCN is larger than IACT of MALA and that the partial updating strategy of MALA-within-Gibbs can further reduce IACT. In particular, we note that the IACT of MALA-within-Gibbs with block size one is nearly identical for the two problem setups, indicating that the dimension independence results we obtained under more restrictive assumptions may indeed hold in practice. As in the previous example, we also note that the step size τ that leads to minimal IACT decreases as we increase the size of the blocks of MALA-within-Gibbs. This is further illustrated in Figure 5.7, where we plot the average acceptance ratio as a function of the step size for MALA-within-Gibbs (several block sizes) and MALA. As in the previous example, we note that for a given, fixed step size, the average acceptance ratio increases as we decrease the block size. The figure also shows IACT as a function of the step size for the various samplers, with optimal step sizes clearly visible. We note, as before, that the partial updating of MALA-within-Gibbs pushes the step size that minimizes IACT towards larger values.

Finally, we note that the acceptance ratio of pCN that minimizes IACT (over the step sizes we tried) in Setup 2 is unusually large (45% rather than about 20%). This is due to insufficient tuning on our part. We considered a range of step sizes and ran pCN chains of length 10^6 for each choice. One can possibly find a step size slightly larger than the 0.01 we tried to reduce the acceptance ratio and decrease IACT. Nonetheless, IACT can be expected to be significantly larger than in the case of MALA or MALA-within-Gibbs. Moreover, given the overall chain length of only 10^6, the estimated IACT of $28,015$ for pCN may not be entirely precise, but all of our numerical experiments indicate that it is in any case very large.

Recall that a small and dimension-independent IACT of MALA-within-Gibbs does not necessarily imply that the algorithm is a computationally efficient sampler (generating one sample requires several likelihood evaluations due to the partial updating strategy, see above). Estimating the cost per effective sample by (5.3), we see that MALA-within-Gibbs is not an efficient sampler for Setup 1, but MALA-within-Gibbs with $q = 68 \times 68$ (leading to two blocks and two likelihood evaluations per sample) is indeed the most effective sampler for Setup 2. This further illustrates that there is a trade-off between the need to reduce IACT by using partial updating and the need to keep the cost-per-sample, which we take to be proportional to the number of blocks, reasonable. In the future, such issues may be addressed by incorporating the partial updating into "local" likelihood evaluations which do not require solving the full PDE, but such issues are beyond the

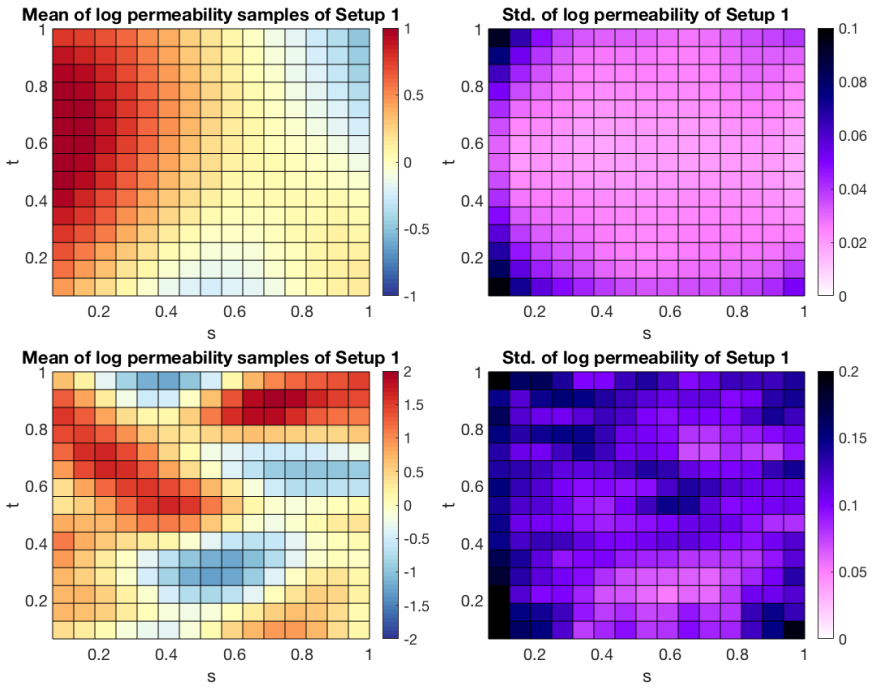

Figure 5.6: Top row: approximate posterior mean (left) and approximate standard deviation (right) computed from $N_e = 10^4$ MALA-within-Gibbs samples with block size $q = 1$ for Setup 1. Bottom row: approximate posterior mean (left) and approximate standard deviation (right) computed from $N_e = 10^3$ MALA-within-Gibbs samples with block size $q = 1$ for Setup 2.

scope of this paper.

We also note that our numerical experiments are limited in the sense that we only considered pCN, MALA, and MALA-within-Gibbs. Other samplers may turn out to be more practical than MALA-within-Gibbs. Specifically, note that the pressure field is relatively well observed, which implies that the posterior differs strongly from the prior (high effective dimension). This explains, at least in part, why we observe such large IACT for pCN. Other "anisotropic" samplers that are modifications of pCN, e.g., DILI [17] or generalized pCN [35], might be effective in this problem. Our goal, however, is not to find the most appropriate sampler for this Bayesian inverse problem, but rather to use this example to demonstrate some of the practical and theoretical aspects of the MALA-within-Gibbs sampler.

6 Conclusion

Markov chain Monte Carlo (MCMC) samplers are used to draw samples from a given target probability distribution in a wide array of applications. We have discussed the numerical efficiency of a particular sampler, the MALA-within-Gibbs sampler, when the target distribution exhibits a particular "sparse conditional structure." In simple terms, the latter is just a structured conditional independence relationship, or block conditional independence relationship, among the variables of interest. For Gaussians, sparse conditional structure is equivalent to a (block-)sparse precision matrix. MALA-within-Gibbs samplers are natural tools to make effective use of sparse conditional structure for numerical efficiency via a suitable partial updating. We have shown that the accep-

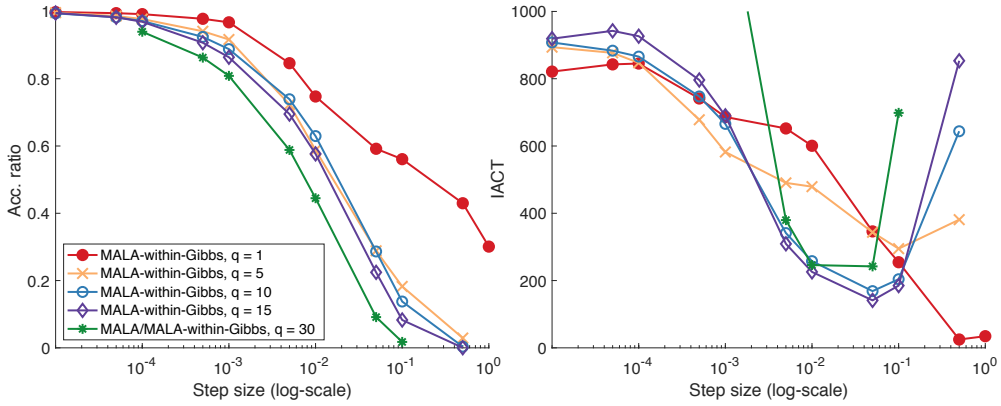

Figure 5.7: Left: average acceptance ratio of MALA-within-Gibbs and MALA, as a function of step size, for Setup 1. Right: average IACT of MALA-within-Gibbs and MALA, as a function of step size, for Setup 1.

tance ratio and step size of MALA-within-Gibbs are independent of the overall dimension of the problem if the partial updating is chosen to be in line with the sparse conditional structure of the target distribution. Under additional assumptions of block-wise log-concavity, we could prove that the convergence rate of MALA-within-Gibbs is independent of dimension. This suggests that MALA-within-Gibbs can be an effective sampler for high-dimensional problems.

We have investigated the applicability of MALA-within-Gibbs in the context of Bayesian inverse problems, where we expect to encounter sparse conditional structure. In many Bayesian inverse problems, we expect that sparse conditional structure can be anticipated based on the prior distribution and the locality of the likelihood, in appropriate coordinates. There are also connections between partial updating in MALA-within-Gibbs and "localization" in numerical weather prediction, which we described briefly. Numerical experiments on two well-known test problems suggest that our theoretical results are indeed indicative of what to expect in practice, where the required assumptions may only hold approximately. For example, in both numerical examples, we could show that measures of performance of the MALA-within-Gibbs sampler, e.g., integrated autocorrelation time (IACT), step size, and acceptance ratio, are indeed independent of the overall dimension of the problem. Nonetheless, the actual computational cost of MALA-within-Gibbs is dependent on the problem dimension because the partial updating requires repeated likelihood evaluations (which are costly) per sample. Our numerical experiments suggest that there is a trade-off between additional computational costs due to the partial updating and the increase in computational cost due to larger IACT or decreasing step size, without partial updating. To keep, for example, IACT small, a large number of partial updates should be used, but this in turn requires several likelihood evaluations per sample. This trade-off may not always be easy to resolve in practice. We have provided examples in which MALA-within-Gibbs leads to significant gains in the computational cost per (effective) sample, but we have also encountered examples in which a global update is, ultimately, the right choice.

References

[1] S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A.M. Stuart. Importance sampling: computational complexity and intrinsic dimension. *Stat. Sci.*, 32(3):405–431, 2017.

[2] M. Asch, M. Bocquet, and M. Nodet. *Data assimilation: methods, algorithms and applications.* SIAM, 2017.

[3] J. Bear. *Modeling groundwater flow and pollution.* Kluwer, 1990.

[4] M. Bédard. Hierarchical models: Local proposal variances for RWM-within-Gibbs and MALA-within-Gibbs. *Comput. Stat. Data Anal.*, 109:231 – 246, 2017.

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

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

[7] Alexandros Beskos. A stable manifold MCMC method for high dimensions. *Statistics & Probability Letters*, 90:46–52, 2014.

[8] T. Bui-Thanh, O. Ghattas, J. Martin, and G. Stadler. A computational framework for infinite-dimensional Bayesian inverse problems. Part I: The linearized case, with application to global seismic inversion. *SIAM J. Sci. Comput.*, 36(4):A2494–A2523, 2013.

[9] N. Chen and A. J. Majda. Filtering nonlinear turbulent dynamical systems through conditional Gaussian statistics. *Mon. Weather Rev.*, 144(12):4885–4917, 2016.

[10] N. Chen and A. J. Majda. Conditional Gaussian systems for multiscale nonlinear stochastic systems: Prediction, state estimation and uncertainty quantification. *Entropy*, 20(7):509, 2018.

[11] N. Chen, A. J. Majda, and X.T. Tong. Spatial localization for nonlinear dynamical stochastic models for excitable media. arXiv:1901.07318.

[12] N. Chen, A. J. Majda, and X.T. Tong. Rigorous analysis for efficient statistically accurate algorithms for solving Fokker-Plank equations in large dimensions. *SIAM-ASA J. Uncertain.*, 6(3):1198–1223, 2018.

[13] Yuxin Chen, David Keyes, Kody JH Law, and Hatem Ltaief. Accelerated dimension-independent adaptive Metropolis. *SIAM Journal on Scientific Computing*, 38(5):S539–S565, 2016.

[14] A.J. Chorin and O.H. Hald. *Stochastic tools in mathematics and science.* Springer, third edition, 2013.

[15] O. F. Christensen, G. O. Roberts, and J. S. Rosenthal. Scaling limits for the transient phase of local Metropolis–Hastings algorithms. *J. R. Stat. Soc. B*, 67(2):253–268, 2005.

[16] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. MCMC methods for functions: modifying old algorithms to make them faster. *Stat. Sci.*, 28(3):424–446, 2013.

[17] T. Cui, K. J. H. Law, and Y. M. Marzouk. Dimension-independent likelihood-informed MCMC. *J. Comput. Phys.*, 304:109–137, 2016.

[18] T. Cui, Y. M. Marzouk, and K. Willcox. Scalable posterior approximations for large-scale Bayesian inverse problems via likelihood-informed parameter and state reduction. *J. Comput. Phys.*, 315:363–387, 2016.

[19] M. Dashti and A. Stuart. Uncertainty quantification and weak approximation of an elliptic inverse problem. *SIAM J. Numer. Anal.*, 49(6):2524–2542, 2011.

[20] C. Fox and A. Parker. Accelerated Gibbs sampling of normal distributions using matrix splittings and polynomials. *Bernoulli*, 23(4B):3711–3743, 2017.

[21] M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. B*, 73:123–214, 2011.

[22] M. Hairer, A.M. Stuart, and S.J. Vollmer. Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *Ann. Appl. Probab.*, 24(6):2455–2490, 2014.

[23] A. A. Johnson, G. L. Jones, and R. C. Neath. Component-wise Markov chain Monte Carlo: Uniform and geometric ergodicity under mixing and composition. *Stat. Sci.*, 28(3):360–375, 08 2013.

[24] Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[25] J. Møller, A. R. Syversveen, and R. P. Waagepetersen. Log Gaussian Cox processes. *Scand. J. Stat.*, 25(3):451–482, 1998.

[26] M. Morzfeld, X.T. Tong, and Y.M. Marzouk. Localization for MCMC: sampling high-dimensional posterior distributions with local structure. *J. Comput. Phys.*, 310:1–28, 2019.

[27] M. Morzfeld, X. Tu, J. Wilkening, and A.J. Chorin. Parameter estimation by implicit sampling. *Comm. App. Math. Com. Sci.*, 10(2):205–225, 2015.

[28] P. Neal and G. O. Roberts. Optimal scaling for partially updating MCMC algorithms. *Ann. Appl. Probab.*, 16(2):475–515, 05 2006.

[29] R. A. Norton and C. Fox. Fast sampling in a linear-Gaussian inverse problem. *SIAM-ASA J. Uncertain.*, 4:1191–1218, 2016.

[30] D. S. Oliver, A. C. Reynolds, and N. Liu. *Inverse theory for petroleum reservoir characterization and history matching*. Cambridge University Press, 2008.

[31] N. Petra, J. Martin, G. Stadler, and O. Ghattas. A computational framework for infinite-dimensional Bayesian inverse problems. Part II: Stochastic Newton MCMC with application to ice sheet flow inverse problems. *SIAM J. Sci. Comput.*, 36(4):A1525–1555, 2013.

[32] S. Reich and C. Cotter. *Probabilistic forecasting and Bayesian data assimilation*. Cambridge University Press, 2015.

[33] G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7:110–120, 1997.

[34] G. O. Roberts and J. S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. B*, 60:255–268, 1998.

[35] Daniel Rudolf and Björn Sprungk. On a generalization of the preconditioned crank–nicolson metropolis algorithm. *Foundations of Computational Mathematics*, 18(2):309–343, 2018.

[36] A. D. Sokal. Monte Carlo methods in statistical mechanics: foundations and new algorithms, 1998.

[37] A. Spantini, D. Bigoni, and Y. M. Marzouk. Inference via low-dimensional couplings. *J. Mach. Learn. Res.*, 19(66):1–71, 2018.

[38] A. Spantini, A. Solonen, T. Cui, J. Martin, L. Tenorio, and Y. M. Marzouk. Optimal low-rank approximations of Bayesian linear inverse problems. *SIAM J. Sci. Comput.*, 37(6):A2451–A2487, 2015.

[39] T.A.Moselhy and Y.M. Marzouk. Bayesian inference with optimal maps. *J. Comput. Phys.*, 231:7815–7850, 2012.

[40] S. Vollmer. Dimension-independent MCMC sampling for inverse problems with non-Gaussian priors. *SIAM-ASA J. Uncertain.*, 3(1):535–561, 2015.

[41] U. Wolff. Monte Carlo errors with less errors. *Comput. Phys. Commun.*, 156(2):143–153, 2004.

[42] Olivier Zahm, Tiangang Cui, Kody Law, Alessio Spantini, and Youssef Marzouk. Certified dimension reduction in nonlinear bayesian inverse problems. *arXiv:1807.03712*, 2018.

A Proofs

In this appendix, we provide the proofs of Proposition 3.3 and Theorem 3.6. The proof strategy relies on a maximal coupling of a pair of MALA-within-Gibbs iterations, say $\mathbf{x}^k$ and $\mathbf{z}^k$. This convergence is independent of the initial distributions of $\mathbf{x}^k$ and $\mathbf{z}^k$, and we pick a $\mathbf{z}^0$ as a sample from $\pi(\cdot)$. Since $\pi(\cdot)$ is the stationary distribution of MALA-within-Gibbs, the distribution of $\mathbf{z}^k$ remains $\pi(\cdot)$ for all k, while $\mathbf{x}^k$ converges to it. Briefly, our proofs consist of four steps.

i) In Section A.1, we discuss coupling of a pair of MALA-within-Gibbs iterations. It will become important to consider how the coupled MALA-within-Gibbs iterations, $\mathbf{x}^k$ and $\mathbf{z}^k$, are accepted or rejected and we distinguish the cases (i) accept $\mathbf{x}^k$ and $\mathbf{z}^k$; (ii) accept $\mathbf{x}^k$, reject $\mathbf{z}^k$; (iii) reject $\mathbf{x}^k$, accept $\mathbf{z}^k$; and (iv) reject $\mathbf{x}^k$ and $\mathbf{z}^k$.

ii) In Section A.2, we derive order τ estimates of the accept/reject probabilities, summarized by Lemma A.2. Proposition 3.3 follows as a corollary.

iii) In Section A.3, we study the dynamics of the "block-update" distance $\|\Delta_j^k\| = \|\mathbf{x}_j^k - \mathbf{z}_j^k\|$.

iv) In Section A.4, it is shown that $\|\Delta_j^k\|$ defines a contraction under the additional assumption of block-wise log-concavity, which implies that the sequence $(\|\Delta_1^k\|, \cdots, \|\Delta_m^k\|)$ converges to zero uniformly. Altogether, this proofs Theorem 3.6.

We will use symbols such as M, M_1, M_2 to denote constants that are independent of the number of blocks, m, or the overall dimension, n. The constants, however, may depend on the dimension of a block, q, the sparsity parameter S and other parameters defined in associated assumptions. The values of the constants M, M_1, M_2 may be different in different places. We re-use M, M_1, and M_2 to avoid introducing many different symbols.

A.1 Coupling block movements

Recall that $\mathbf{x}^k$ denotes iterates of the MALA-within-Gibbs algorithm, with $\mathbf{x}^0$ sampled from a certain initial distribution. Now we consider another sequence of iterations of MALA-within-Gibbs,

denoted by $\mathbf{z}^k$. The initial distribution of $\mathbf{z}^0$ is set to be the target distribution $\pi(\cdot)$. Since $\pi(\cdot)$ is the stationary distribution of MALA-within-Gibbs, the distribution of $\mathbf{z}^k$ is π for all k.

To discuss the block updates within each Gibbs iteration, we use

$$\mathbf{x}^{k,j} = [\mathbf{x}_1^{k+1}, \cdots, \mathbf{x}_{j-1}^{k+1}, \mathbf{x}_j^k, \cdots, \mathbf{x}_m^k], \quad \mathbf{z}^{k,j} = [\mathbf{z}_1^{k+1}, \cdots, \mathbf{z}_{j-1}^{k+1}, \mathbf{z}_j^k, \cdots, \mathbf{z}_m^k],$$

to denote the state of the k-th Gibbs cycle before the MALA update of the j-th block. With this notation, the i-th block of $\mathbf{x}^{k,j}$, denoted by $\mathbf{x}_i^{k,j}$, is $\mathbf{x}_i^{k+1}$ if $i < j$ and is $\mathbf{x}_i^k$ if $i \geq j$.

The j-th block proposal made to $\mathbf{x}^{k,j}$ is given by (2.1), and likewise for $\mathbf{z}^{k,j}$. We consider coupling the random noises in the two proposals, so they share the same ξ_j^k. In other words, the proposals are

$$\tilde{\mathbf{x}}_j^k = \mathbf{x}_j^k + \tau \mathbf{v}_j(\mathbf{x}^{k,j}) + \sqrt{2\tau}\xi_j^k, \quad \tilde{\mathbf{z}}_j^k = \mathbf{z}_j^k + \tau \mathbf{v}_j(\mathbf{z}^{k,j}) + \sqrt{2\tau}\xi_j^k.$$

We combine them with other blocks from $\mathbf{x}^{k,j}$ and $\mathbf{z}^{k,j}$, and define

$$\tilde{\mathbf{x}}^{k,j} := [\mathbf{x}_1^{k+1}, \cdots, \mathbf{x}_{j-1}^{k+1}, \tilde{\mathbf{x}}_j^k, \mathbf{x}_{j+1}^k, \cdots, \mathbf{x}_m^k], \quad \tilde{\mathbf{z}}^{k,j} := [\mathbf{z}_1^{k+1}, \cdots, \mathbf{z}_{j-1}^{k+1}, \tilde{\mathbf{z}}_j^k, \mathbf{z}_{j+1}^k, \cdots, \mathbf{z}_m^k].$$

The probabilities with which the proposals $\tilde{\mathbf{x}}^{k,j}$ and $\tilde{\mathbf{z}}^{k,j}$ are accepted are $\alpha_j(\mathbf{x}^{k,j}, \tilde{\mathbf{x}}^{k,j})$ and $\alpha_j(\mathbf{z}^{k,j}, \tilde{\mathbf{z}}^{k,j})$ respectively. The accept/reject step is equivalent to comparing $\alpha_j(\cdot)$ with a random variable, uniformly distributed over $[0, 1]$. Specifically, let $U_{j,\mathbf{x}}^k$ be a draw from a uniform distirbution. The proposal $\tilde{\mathbf{x}}^{k,j}$ is accepted if $U_{j,\mathbf{x}}^k \leq \alpha_j(\mathbf{x}^{k,j}, \tilde{\mathbf{x}}^{k,j})$. Similarly, the proposal $\tilde{\mathbf{z}}_j^k$ is accepted if $U_{j,\mathbf{z}}^k \leq \alpha_j(\mathbf{z}^{k,j}, \tilde{\mathbf{z}}^{k,j})$, where $U_{j,\mathbf{z}}^k$ is a draw from a uniform distribution. A maximal coupling of the acceptance steps is achieved by setting $U_{j,\mathbf{x}}^k = U_{j,\mathbf{z}}^k = U_j^k$. More specifically, there are four scenarios for the acceptance, based on the value of U_j^k:

$$\begin{cases} \text{Both accept} & \text{if } U_j^k \leq \alpha_j(\mathbf{x}^{k,j}, \tilde{\mathbf{x}}^{k,j}) \wedge \alpha_j(\mathbf{z}^{k,j}, \tilde{\mathbf{z}}^{k,j}), \\ \text{Both reject} & \text{if } \alpha_j(\mathbf{x}^{k,j}, \tilde{\mathbf{x}}^{k,j}) \vee \alpha_j(\mathbf{z}^{k,j}, \tilde{\mathbf{z}}^{k,j}) < U_j^k, \\ \text{Accept } \tilde{\mathbf{z}} \text{ reject } \tilde{\mathbf{x}} & \text{if } \alpha_j(\mathbf{x}^{k,j}, \tilde{\mathbf{x}}^{k,j}) \wedge \alpha_j(\mathbf{z}^{k,j}, \tilde{\mathbf{z}}^{k,j}) < U_j^k \leq \alpha_j(\mathbf{z}^{k,j}, \tilde{\mathbf{z}}^{k,j}), \\ \text{Accept } \tilde{\mathbf{x}} \text{ reject } \tilde{\mathbf{z}} & \text{if } \alpha_j(\mathbf{x}^{k,j}, \tilde{\mathbf{x}}^{k,j}) \wedge \alpha_j(\mathbf{z}^{k,j}, \tilde{\mathbf{z}}^{k,j}) < U_j^n \leq \alpha_j(\mathbf{x}^{k,j}, \tilde{\mathbf{x}}^{k,j}). \end{cases} \tag{1.1}$$

Here, "accept" means to set $\mathbf{x}_j^{k+1} = \tilde{\mathbf{x}}_j^k$ and "reject" means to set $\mathbf{x}_j^{k+1} = \mathbf{x}_j^k$, and likewise for $\tilde{\mathbf{z}}$; moreover, $a \wedge b := \min\{a, b\}$, and $a \vee b := \max\{a, b\}$. It is straightforward to verify that marginally $\mathbf{x}_j^{k+1}$ and $\mathbf{z}_j^{k+1}$ follow the same distribution (as described in the MALA-within-Gibbs algorithm).

The information before the update of $\mathbf{x}_j^k, \mathbf{z}_j^k$ is given by the filtration

$$\mathcal{F}_{k,j} = \sigma\{\mathbf{z}^0, \mathbf{x}^0, \xi_i^{t-1}, U_i^{t-1}, \xi_s^t, U_s^t, t \leq k-1, s \leq j-1, i = 1, \ldots, m\}.$$

For simplicity we write $\mathcal{F}_k := \mathcal{F}_{k,1}$, which is the information available when the k-th Gibbs cycle starts. It is clear that $\mathbf{x}^k, \mathbf{z}^k \in \mathcal{F}_k$. We denote the conditional expectation (probability) w.r.t. $\mathcal{F}_{k,j}$ and $\mathcal{F}_k$ as $\mathbb{E}_{k,j}(\mathbb{P}_{k,j})$ and $\mathbb{E}_k(\mathbb{P}_k)$ respectively.

A.2 Block-acceptance probabilities (proof or Proposition 3.3)

We first derive order estimates of the accept/reject probabilities, with respect to τ, by calculating the derivatives of the acceptance probability. We do so by establishing a Lemma for a function f such that

$$\alpha_j(\mathbf{x}^{k,j}, \tilde{\mathbf{x}}^{k,j}) = f(\mathbf{x}^{k,j}, \sqrt{\tau}, \xi_j^k) \wedge 1,$$

where $\alpha_j(\mathbf{x}^{k,j}, \tilde{\mathbf{x}}^{k,j})$ is the acceptance probability of the jth block in a MALA-within-Gibbs iteration. The Lemma is the used to prove Proposition 3.3. We will make repeated use of the fact that $\frac{\nabla_{\mathbf{x}}\pi(\mathbf{x})}{\pi(\mathbf{x})} = \nabla_{\mathbf{x}} \log \pi(\mathbf{x}) = \mathbf{v}(\mathbf{x})$. We simplify the notation by writing $w = \sqrt{\tau}$.

Lemma A.1. *Suppose Assumptions 3.1 and 3.2 hold. Fix a $j \in \{1, \cdots, m\}$ and, for any given $\mathbf{x} \in \mathbb{R}^n, \xi \in \mathbb{R}^{q_j}, q_j = dim(\mathbf{x}_j), w \in [0,1]$, define*

$$f(\mathbf{x}, w, \xi) := \frac{\pi(\tilde{\mathbf{x}}) \exp(-\frac{1}{4w^2}\|\mathbf{x}_j - \tilde{\mathbf{x}}_j - w^2\mathbf{v}_j(\tilde{\mathbf{x}})\|^2)}{\pi(\mathbf{x}) \exp(-\frac{1}{4w^2}\|\tilde{\mathbf{x}}_j - \mathbf{x}_j - w^2\mathbf{v}_j(\mathbf{x})\|^2)}.$$

Here the blocks of $\tilde{\mathbf{x}}$ are given by $\tilde{\mathbf{x}}_j = \mathbf{x}_j + w^2\mathbf{v}_j(\mathbf{x}) + \sqrt{2}w\xi$ (see Equation (2.1)), and $\tilde{\mathbf{x}}_i = \mathbf{x}_i$ for $i \neq j$. Then

(1) If $i \notin \mathcal{I}_j$, $\nabla_{\mathbf{x}_i} f(\mathbf{x}, w, \xi) = \mathbf{0}$.

(2) There is a constant M such that $\|\nabla_{\mathbf{x}} f(\mathbf{x}, w, \xi)\| \leq w^2 M(\|\xi\|^2 + 1) f(\mathbf{x}, w, \xi)$.

(3) There is a constant M such that $|\partial_w f(\mathbf{x}, w, \xi)| \leq M(\|\xi\|^2 + 1) f(\mathbf{x}, w, \xi)$.

Proof. Note that f can be rewritten as

$$\begin{aligned} f &= \frac{\pi(\tilde{\mathbf{x}})}{\pi(\mathbf{x})} \exp\left(-\frac{1}{4}\|w\mathbf{v}_j(\mathbf{x}) + w\mathbf{v}_j(\tilde{\mathbf{x}}) + \sqrt{2}\xi\|^2 + \frac{1}{2}\|\xi\|^2\right) \\ &= \frac{\pi(\tilde{\mathbf{x}})}{\pi(\mathbf{x})} \exp\left(-\frac{1}{4}(w^2\|\mathbf{v}_j(\mathbf{x})\|^2 + w^2\|\mathbf{v}_j(\tilde{\mathbf{x}})\|^2 + 2w^2\mathbf{v}_j(\mathbf{x})^T\mathbf{v}_j(\tilde{\mathbf{x}}) + 2\sqrt{2}w\mathbf{v}_j(\mathbf{x})^T\xi + 2\sqrt{2}w\mathbf{v}_j(\tilde{\mathbf{x}})^T\xi)\right), \end{aligned}$$

where $\mathbf{v}_j$ is a q_j-dimensional vector. The fact that the dimension of $\mathbf{v}_j$ is q_j, rather than n, makes derivations cumbersome. We thus pad $\mathbf{v}_j$ with zero blocks to form an n-dimensional vector $\tilde{\mathbf{v}}_j$, where the j-th block of $\tilde{\mathbf{v}}_j$ is equal to $\mathbf{v}_j(\mathbf{x})$, but all other components are zero. Similarly, we form an n-dimensional $\tilde{\xi}$ from a q_j-dimensional ξ so that the jth block of $\tilde{\xi}$ is equal to ξ, but all other components are zero. With this notation,

$$\tilde{\mathbf{x}} = \mathbf{x} + w^2\tilde{\mathbf{v}}_j(\mathbf{x}) + \sqrt{2}w\tilde{\xi}.$$

The associated Jacobians are given by

$$\nabla_{\mathbf{x}}\tilde{\mathbf{x}} = \mathbf{I} + w^2\nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x}), \quad \nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}}) = \nabla_{\tilde{\mathbf{x}}}\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}})(\mathbf{I} + w^2\nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x}))$$

Since, by construction, $\|\mathbf{v}_j\|^2 = \|\tilde{\mathbf{v}}_j\|^2$, we have

$$\begin{aligned} \log f = &\log\pi(\tilde{\mathbf{x}}) - \log\pi(\mathbf{x}) \\ &- \frac{1}{4}(w^2\|\tilde{\mathbf{v}}_j(\mathbf{x})\|^2 + w^2\|\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}})\|^2 + 2w^2\tilde{\mathbf{v}}_j(\mathbf{x})^T\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}}) + 2\sqrt{2}w\tilde{\mathbf{v}}_j(\mathbf{x})^T\tilde{\xi} + 2\sqrt{2}w\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}})^T\tilde{\xi}). \end{aligned}$$

The chain rule then gives the gradient of f:

$$\begin{aligned} \frac{\nabla_{\mathbf{x}} f}{f} = \nabla_{\mathbf{x}} \log f = &(\mathbf{I} + w^2\nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x}))^T\mathbf{v}(\tilde{\mathbf{x}}) - \mathbf{v}(\mathbf{x}) - \frac{1}{2}w^2(\nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x}))^T\tilde{\mathbf{v}}_j(\mathbf{x}) \\ &- \frac{1}{2}w^2(\mathbf{I} + w^2\nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x}))^T(\nabla_{\tilde{\mathbf{x}}}\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}}))^T\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}}) \\ &- \frac{1}{2}w^2\nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x})^T\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}}) - \frac{1}{2}w^2(\mathbf{I} + w^2\nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x}))^T\nabla_{\tilde{\mathbf{x}}}\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}})^T\tilde{\mathbf{v}}_j(\mathbf{x}) \\ &- \frac{1}{2}\sqrt{2}w(\nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x})^T + (\mathbf{I} + w^2\nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x}))^T\nabla_{\tilde{\mathbf{x}}}\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}})^T)\tilde{\xi}. \end{aligned} \tag{1.2}$$

We will first verify claim (1). Let $\mathbf{u}$ be a vector that has nonzero components only in the i-th block. Then, claim (1) is equivalent to showing $\mathbf{u}^T(\nabla_{\mathbf{x}} \log f) = 0$. Note the only non-zero blocks

of $\nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x})$ are in its j-th row, so that only the jth block of $\mathbf{u}^T\nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x})^T$ is nonzero. This block can be written as $\mathbf{u}_i^T\nabla_{\mathbf{x}_i}\tilde{\mathbf{v}}_j(\mathbf{x})^T$. By Assumption 3.2, and since $i \notin \mathcal{I}_j$,

$$\nabla_{\mathbf{x}_i}\tilde{\mathbf{v}}_j(\mathbf{x}) = \nabla_{\mathbf{x}_i}\nabla_{\mathbf{x}_j}\log\pi = \mathbf{0} \quad \Rightarrow \quad \mathbf{u}^T\nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x})^T = \mathbf{0}^T.$$

Knowing this can simplify the computation of $\mathbf{u}^T(\nabla_{\mathbf{x}}\log f)$, since most terms in (1.2) include the Jacobian matrix, and they drop out when multiplying with $\mathbf{u}^T$. The gradient of f in (1.2) now simplifies to $\mathbf{u}^T(\nabla_{\mathbf{x}}\log f) = \langle \mathbf{u}, \mathbf{v}(\tilde{\mathbf{x}}) - \mathbf{v}(\mathbf{x})\rangle$, where we use $\langle a, b\rangle = a^T b$ to denote an inner product. Note that $\tilde{\mathbf{x}}$ differs from $\mathbf{x}$ only in its jth block. Writing $\Delta = \tilde{\mathbf{x}} - \mathbf{x} \in \mathbb{R}^{qm}$, we obtain

$$\langle \mathbf{u}, \mathbf{v}(\tilde{\mathbf{x}}) - \mathbf{v}(\mathbf{x})\rangle = \left\langle \mathbf{u}, \int_0^1 \nabla_{\mathbf{x}}\mathbf{v}(\mathbf{x} + s\Delta)ds\Delta \right\rangle = 0,$$

because $\mathbf{u}$ has nonzero entries only outside the j-th block, and Δ has nonzero entries only in j-th block, and by Assumption 3.2, the $(\mathbf{x}_i, \mathbf{x}_j)$-th block of $\nabla_{\mathbf{x}}\mathbf{v} = \nabla_{\mathbf{x}}^2\log\pi$ is zero.

For claim (2), we collect terms of the same w order in (1.2):

$$\frac{\nabla_{\mathbf{x}} f}{f} = (\mathbf{v}(\tilde{\mathbf{x}}) - \mathbf{v}(\mathbf{x})) - \frac{1}{2}\sqrt{2}w(\nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x}) + \nabla_{\tilde{\mathbf{x}}}\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}}))^T\tilde{\xi} + w^2 R(\xi, \mathbf{x}, w).$$

By Assumption 3.2, the elements of all vectors and matrices appearing above are bounded, so the residual term can be bounded by $\|R(\xi, \mathbf{x}, w)\| \le (\|\xi\| + 1)M_1$ with a constant M_1. By adding and then subtracting a term, we have the following bound

$$\begin{aligned}\left\|\frac{\nabla_{\mathbf{x}} f}{f}\right\| \le& \|\mathbf{v}(\tilde{\mathbf{x}}) - \mathbf{v}(\mathbf{x}) - \nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x})^T(\tilde{\mathbf{x}} - \mathbf{x})\| + w^2(\|\xi\| + 1)M_1 \\ &+ \left\|\frac{1}{2}\sqrt{2}w(\nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x}) + \nabla_{\tilde{\mathbf{x}}}\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}}))^T\tilde{\xi} - \nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x})^T(\tilde{\mathbf{x}} - \mathbf{x})\right\|. \end{aligned} \tag{1.3}$$

To bound the first term on the right hand side of (1.3), first note that because the j-th row block of $\nabla_{\mathbf{x}}\mathbf{v}(\mathbf{x})$ and $\nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x})$ are the same, and $(\tilde{\mathbf{x}} - \mathbf{x})$ has nonzero entries only in the j-th block,

$$\nabla_{\mathbf{x}}\mathbf{v}(\mathbf{x})(\tilde{\mathbf{x}} - \mathbf{x}) = \nabla_{\mathbf{x}}\mathbf{v}(\mathbf{x})^T(\tilde{\mathbf{x}} - \mathbf{x}) = \nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x})^T(\tilde{\mathbf{x}} - \mathbf{x}),$$

where the first identity is due to the fact that the Hessian of $\log\pi$, $\nabla_{\mathbf{x}}\mathbf{v}(\mathbf{x})$, is symmetric. The first term in (1.3) can therefore be bounded by a Taylor expansion of $\mathbf{v}$, followed by Assumption 3.2 and Cauchy inequality,

$$\begin{aligned}&\|\mathbf{v}(\tilde{\mathbf{x}}) - \mathbf{v}(\mathbf{x}) - \nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x})^T(\tilde{\mathbf{x}} - \mathbf{x})\| = \|\mathbf{v}(\tilde{\mathbf{x}}) - \mathbf{v}(\mathbf{x}) - \nabla_{\mathbf{x}}\mathbf{v}(\mathbf{x})(\tilde{\mathbf{x}} - \mathbf{x})\| \\ &\le H_v\|\tilde{\mathbf{x}} - \mathbf{x}\|^2 = H_v\|\tilde{\mathbf{x}}_j - \mathbf{x}_j\|^2 = H_v\|w^2\mathbf{v}_j(\mathbf{x}) + \sqrt{2}w\xi\|^2 \le (4w^2\|\xi\|^2 + 2w^4M_v^2)H_v. \end{aligned} \tag{1.4}$$

To bound the second term on the right hand side of (1.3), note that Assumption 3.2, combined with the Taylor expansion and Young's inequality, gives

$$\|\nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x}) - \nabla_{\tilde{\mathbf{x}}}\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}})\| \le H_v\|\mathbf{x} - \tilde{\mathbf{x}}\| \le (\sqrt{2}w\|\xi\| + w^2M_v)H_v.$$

Recall that $\tilde{\xi}$ is equal to ξ, padded with zeros so that $\|\tilde{\xi}\| = \|\xi\|$. Thus, the second term on the

right hand side of (1.3) is bounded by

$$\begin{aligned}
&\left\| \frac{1}{2}\sqrt{2}w(\nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x}) + \nabla_{\tilde{\mathbf{x}}}\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}}))^T\tilde{\xi} - \nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x})^T(\tilde{\mathbf{x}} - \mathbf{x}) \right\| \\
&\leq \left\| \sqrt{2}w\nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x})^T\tilde{\xi} - \nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x})^T(\tilde{\mathbf{x}} - \mathbf{x}) \right\| + \left\| \frac{1}{2}\sqrt{2}w(\nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x}) - \nabla_{\tilde{\mathbf{x}}}\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}}))^T\tilde{\xi} \right\| \\
&\leq \left\| \sqrt{2}w\nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x})^T\tilde{\xi} - \nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x})^T(\tilde{\mathbf{x}} - \mathbf{x}) \right\| + (w^2\|\xi\|^2 + w^3\|\xi\|M_v)H_v \\
&\leq \left\| \sqrt{2}w\nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x})^T\tilde{\xi} - \nabla_{\mathbf{x}}\tilde{\mathbf{v}}_j(\mathbf{x})^T(\sqrt{2}w\tilde{\xi} + w^2\tilde{\mathbf{v}}_j(\mathbf{x})) \right\| + (w^2\|\xi\|^2 + w^3\|\xi\|M_v)H_v \\
&\leq w^2M_vH_v + w^2H_v\|\xi\|^2 + w^3\|\xi\|M_vH_v. \qquad (1.5)
\end{aligned}$$

If we replace the terms in (1.3) with the bounds in (1.4) and (1.5), and if we use the fact that $w \leq 1$, we find that there exists a constant M such that

$$\left\| \frac{\nabla_{\mathbf{x}} f}{f} \right\| \leq w^2M(\|\xi\|^2 + 1),$$

which is our claim (2).

For claim (3), consider the derivative of $\tilde{\mathbf{x}}$ with respect to w

$$\|\partial_w\tilde{\mathbf{x}}\| = \|2w\tilde{\mathbf{v}}_j(\mathbf{x}) + \sqrt{2}\tilde{\xi}\| \leq \sqrt{2}\|\xi\| + 2wM_v,$$

using, again, that $w \leq 1$. Moreover, since $\nabla_{\tilde{\mathbf{x}}}\tilde{\mathbf{v}}_j$ has nonzero blocks only on the j-th row, we have that

$$\|\nabla_{\tilde{\mathbf{x}}}\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}})\| \leq \sum_{i\in\mathcal{I}_j} \|\nabla_{\tilde{\mathbf{x}}_i}\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}})\| \leq SH_v.$$

Therefore,

$$\|\partial_w\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}})\| = \|\nabla_{\tilde{\mathbf{x}}}\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}})\partial_w\tilde{\mathbf{x}}\| \leq SH_v(\sqrt{2}\|\xi\| + 2wM_v).$$

Finally, recall that

$$\log f = \log\pi(\tilde{\mathbf{x}}) - \log\pi(\mathbf{x}) - \frac{1}{4}(w^2\|\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}}) + \tilde{\mathbf{v}}_j(\mathbf{x})\|^2 + 2\sqrt{2}w\tilde{\xi}^T(\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}}) + \tilde{\mathbf{v}}_j(\mathbf{x}))),$$

so that

$$\begin{aligned}
\frac{\partial_w f(\mathbf{x}, w, \xi)}{f(\mathbf{x}, w, \xi)} &= \left\langle \mathbf{v}(\tilde{\mathbf{x}}), 2w\tilde{\mathbf{v}}_j(\mathbf{x}) + \sqrt{2}\tilde{\xi} \right\rangle - \frac{1}{2}w\|\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}}) + \tilde{\mathbf{v}}_j(\mathbf{x})\|^2 \\
&\quad - \frac{1}{2}w^2(\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}}) + \tilde{\mathbf{v}}_j(\mathbf{x}))^T\partial_w\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}}) - \frac{\sqrt{2}}{2}\tilde{\xi}^T(\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}}) + \tilde{\mathbf{v}}_j(\mathbf{x})) - \frac{\sqrt{2}}{2}w\tilde{\xi}^T\partial_w\tilde{\mathbf{v}}_j(\tilde{\mathbf{x}}).
\end{aligned}$$

Further, recall that the padding with zeros does not affect the inner product:

$$\left\langle \mathbf{v}(\tilde{\mathbf{x}}), 2w\tilde{\mathbf{v}}_j(\mathbf{x}) + \sqrt{2}\tilde{\xi} \right\rangle = \left\langle \tilde{\mathbf{v}}_j(\tilde{\mathbf{x}}), 2w\tilde{\mathbf{v}}_j(\mathbf{x}) + \sqrt{2}\tilde{\xi} \right\rangle.$$

Since every term in the expression of $\frac{\partial_w f(\mathbf{x},w,\xi)}{f(\mathbf{x},w,\xi)}$ is, by Assumption 3.2, bounded, we can conclude there exists an M such that

$$|\partial_w f(\mathbf{x}, w, \xi)| \leq (M + M\|\xi\|^2)f(\mathbf{x}, w, \xi).$$

This leads us to claim (3). □

Lemma A.2. *Under Assumptions 3.1 and 3.2, there exists a constant M, such that, for any $\mathbf{x}^{k,j}$ and $\mathbf{z}^{k,j}$, coupled by one step of MALA-within-Gibbs algorithm at block j as in* (1.1)*, we have that*

(1) $\mathbb{P}_{k,j}(\textit{accept both}) \geq 1 - M\sqrt{\tau}$,

(2) $\mathbb{P}_{k,j}(\textit{accept only one}) \leq M\tau\sqrt{\sum_{i\in\mathcal{I}_j}\|\mathbf{x}_i^{k,j} - \mathbf{z}_i^{k,j}\|^2}$.

(3) $\mathbb{E}_{k,j}\|\xi_j^k\|\mathbb{1}_{\textit{accept only one}} \leq M\tau\sqrt{\sum_{i\in\mathcal{I}_j}\|\mathbf{x}_i^{k,j} - \mathbf{z}_i^{k,j}\|^2}$.

Proposition 3.3 follows immediately by the Tower property since

$$\mathbb{E}[\alpha_j(\mathbf{x}^k, \tilde{\mathbf{x}}^k)] = \mathbb{E}[\mathbb{E}_{k,j}\alpha_j(\mathbf{x}^k, \tilde{\mathbf{x}}^k)] \geq \mathbb{E}[\mathbb{P}_{k,j}(\textit{accept both})].$$

Proof. Since here we are concerned with updating one block, for simplicity of the notation, we write $\mathbf{x} = \mathbf{x}^{k,j}, \tilde{\mathbf{x}} = \tilde{\mathbf{x}}^{k,j}, \mathbf{z} = \mathbf{z}^{k,j}, \tilde{\mathbf{z}} = \tilde{\mathbf{z}}^{k,j}$. Let "reject $\tilde{\mathbf{x}}$" denote the event that the proposal of $\tilde{\mathbf{x}}$ is rejected. We will show below that $\mathbb{P}_{k,j}(\text{reject } \tilde{\mathbf{x}}) \leq \frac{1}{2}M\sqrt{\tau}$ for a certain M. Thus,

$$\mathbb{P}_{k,j}(\text{accept both}) \geq 1 - \mathbb{P}_{k,j}(\text{reject } \tilde{\mathbf{x}}) - \mathbb{P}_{k,j}(\text{reject } \tilde{\mathbf{z}}) \geq 1 - M\sqrt{\tau}.$$

Let $f(\mathbf{x}, w, \xi)$ be as defined by Lemma A.1. Note that

$$\mathbb{P}_{k,j}(\text{reject } \tilde{\mathbf{x}}) = \mathbb{E}_{k,j}(1 - f(\mathbf{x}, \sqrt{\tau}, \xi_j^k) \wedge 1).$$

Next we bound $1 - f(\mathbf{x}, \sqrt{\tau}, \xi_j^k) \wedge 1$. Note that $f(\mathbf{x}, 0, \xi_j^k) = 1$. For each fixed $\mathbf{x}, \xi_j^k$, if $f(\mathbf{x}, y, \xi_j^k) \leq 1$ for all $y \in (0, \sqrt{\tau}]$, let $w_1 = w_2 = \sqrt{\tau}$; otherwise, let

$$w_1 = \inf\{y \in [0, \sqrt{\tau}] : f(\mathbf{x}, y, \xi_j^k) > 1\}, \quad w_2 = \sup\{y \in [w_1, \sqrt{\tau}] : f(\mathbf{x}, y, \xi_j^k) \geq 1\}.$$

One can check that the following holds with either $f(\mathbf{x}, \sqrt{\tau}, \xi_j^k) < 1$ or $f(\mathbf{x}, \sqrt{\tau}, \xi_j^k) \geq 1$

$$1 \wedge f(\mathbf{x}, \sqrt{\tau}, \xi_j^k) - f(\mathbf{x}, 0, \xi_j^k) = \int_{w_2}^{\sqrt{\tau}} \partial_y f(\mathbf{x}, y, \xi_j^k)dy + \int_0^{w_1} \partial_y f(\mathbf{x}, y, \xi_j^k)dy. \tag{1.6}$$

Note also that by the definition of w_1 and w_2,

$$f(\mathbf{x}, y, \xi_j^k) \leq 1, \quad \forall y \in [0, w_1] \cup [w_2, \sqrt{\tau}].$$

Thus, by Lemma A.1 (claim (3)), there is a constant M_1 so that:

$$|\partial_y f(\mathbf{x}, y, \xi_j^k)| \leq (M_1 + M_1\|\xi_j^k\|^2)|f(\mathbf{x}, y, \xi_j^k)| \leq M_1 + M_1\|\xi_j^k\|^2, \quad y \in [0, w_1] \cup [w_2, \sqrt{\tau}].$$

Applying this upper bound to the integrant in (1.6), and using the fact that $f(\mathbf{x}, 0, \xi_j^k) = 1$, we obtain:

$$1 - 1 \wedge f(\mathbf{x}, \sqrt{\tau}, \xi_j^k) \leq \sqrt{\tau}M_1(1 + \|\xi_j^k\|^2).$$

Recall that ξ_j^k is a sample of a standard normal variable whose dimension is less than q by Assumption 3.1 (q being the dimension of one block). Consequentially, $\mathbb{E}_{k,j}(1 - f(\mathbf{x}, \sqrt{\tau}, \xi_j^k) \wedge 1) \leq \sqrt{\tau}M_1(1 + q)$, which leads to our first claim.

As for the second claim, given $\mathbf{x}, \mathbf{z}, \xi_j^k$, the probability of having only one proposal being accepted is $|f(\mathbf{x}, \sqrt{\tau}, \xi_j^k) \wedge 1 - f(\mathbf{z}, \sqrt{\tau}, \xi_j^k) \wedge 1|$. Let $\mathbf{y}_s$ be the linear interpolation between $\mathbf{x}$ and $\mathbf{z}$, so that $\mathbf{y}_0 = \mathbf{x}$ and $\mathbf{y}_1 = \mathbf{z}$. If $f(\mathbf{x}, \sqrt{\tau}, \xi_j^k)$ and $f(\mathbf{z}, \sqrt{\tau}, \xi_j^k)$ are both above 1, then our claim holds trivially.

Otherwise, either $f(\mathbf{x}, \sqrt{\tau}, \xi_j^k)$ or $f(\mathbf{z}, \sqrt{\tau}, \xi_j^k)$ is less than 1. Assume, without loss of generality, that $f(\mathbf{x}, \sqrt{\tau}, \xi_j^k) < 1$ and define

$$s^* = \begin{cases} \inf\{s \in [0,1] : f(\mathbf{y}_s, \sqrt{\tau}, \xi_j^k) \geq 1\}, & f(\mathbf{z}, \sqrt{\tau}, \xi_j^k) > 1; \\ 1, & \text{else.} \end{cases}$$

Then, for $s \in [0, s^*]$, $f(\mathbf{y}_s, \sqrt{\tau}, \xi_j^k) \leq 1$. Also note that, by Lemma A.1 (claim(1)),

$$f(\mathbf{x}, \sqrt{\tau}, \xi_j^k) = f([\mathbf{x}_1, \dots \mathbf{x}_m], \sqrt{\tau}, \xi_j^k)$$

has dependence on $\mathbf{x}_i$ only if $i \in \mathcal{I}_j$. Therefore, by Lemma A.1 (2), there is a constant M_2, such that

$$\begin{aligned} |f(\mathbf{x}, \sqrt{\tau}, \xi_j^k) \wedge 1 - f(\mathbf{z}, \sqrt{\tau}, \xi_j^k) \wedge 1| &\leq |f(\mathbf{y}_0, \sqrt{\tau}, \xi_j^k) - f(\mathbf{y}_{s^*}, \sqrt{\tau}, \xi_j^k)| \\ &= \left| \int_0^{s_*} \nabla_{\mathbf{y}_s} f(\mathbf{y}_s, \sqrt{\tau}, \xi_j^k)(\mathbf{x}^{k,j} - \mathbf{z}^{k,j}) ds \right| \\ &\leq \sqrt{\sum_{i \in \mathcal{I}_j} \|\mathbf{x}_i^{k,j} - \mathbf{z}_i^{k,j}\|^2} \int_0^{s_*} \|\nabla_{\mathbf{y}_s} f(\mathbf{y}_s, \sqrt{\tau}, \xi_j^k)\| ds \\ &\leq \sqrt{\sum_{i \in \mathcal{I}_j} \|\mathbf{x}_i^{k,j} - \mathbf{z}_i^{k,j}\|^2} \int_0^1 \frac{\|\nabla_{\mathbf{y}_s} f(\mathbf{y}_s, \sqrt{\tau}, \xi_j^k)\|}{f(\mathbf{y}_s)} ds \\ &\leq \sqrt{\sum_{i \in \mathcal{I}_j} \|\mathbf{x}_i^{k,j} - \mathbf{z}_i^{k,j}\|^2} M_2(\|\xi_j^k\|^2 + 1)\tau. \end{aligned}$$

Averaging the above expression over all possible outcomes of ξ_j^k proves our second claim.

Similarly, for the third claim, we have

$$\begin{aligned} &\mathbb{E}_{k,j} |f(\mathbf{x}, \sqrt{\tau}, \xi_j^k) \wedge 1 - f(\mathbf{z}, \sqrt{\tau}, \xi_j^k) \wedge 1| \|\xi_j^k\| \\ &\leq \mathbb{E}_{k,j} \sqrt{\sum_{i \in \mathcal{I}_j} \|\mathbf{x}_i^{k,j} - \mathbf{z}_i^{k,j}\|^2} M_2(\|\xi_j^k\|^3 + \|\xi_j^k\|)\tau. \end{aligned}$$

The upper bound in claim (3) can be obtained by averaging over all ξ_j^k. □

A.3 Block-distance updates

We focus on the difference between the pair of coupled MALA-within-Gibbs iterations and define

$$\Delta_j^k := \mathbf{x}_j^k - \mathbf{z}_j^k, \quad \Delta^{k,j} = \mathbf{x}^{k,j} - \mathbf{z}^{k,j}, \tag{1.7}$$

where each block of $\Delta^{k,j}$ is given by

$$\Delta_i^{k,j} = \mathbf{x}_i^{k,j} - \mathbf{z}_i^{k,j} = \begin{cases} \Delta_i^{k+1}, & i \leq j-1; \\ \Delta_i^k, & i \geq j. \end{cases}$$

We first analyze the dynamics of $\|\Delta_j^k\|$ and use the results to prove Theorem 3.6 under additional assumptions of block-wise log-concavity.

Proposition A.3. *Under Assumptions 3.1 and 3.2, there is a constant M such that after the j-th MALA-within-Gibbs step at the k-th iteration,*

$$\mathbb{E}_{k,j}(\|\Delta_j^{k+1}\|) \le \|\Delta_j^k + \tau \mathbf{v}_j(\mathbf{x}^{k,j}) - \tau \mathbf{v}_j(\mathbf{z}^{k,j})\| + M\tau^{\frac{3}{2}} \sum_{i \in \mathcal{I}_j} \|\Delta_i^{k,j}\|. \tag{1.8}$$

Proof. Recall that the coupling of the acceptance steps has four scenarios (both accepted, both rejected, one rejected the other accepted). If both proposals are rejected, then

$$\|\Delta_j^{k+1}\| = \|\Delta_j^k\|.$$

If both proposals are accepted, then

$$\|\Delta_j^{k+1}\| = \|\mathbf{x}_j^k + \tau \mathbf{v}_j(\mathbf{x}^{k,j}) - \mathbf{z}_j^k - \tau \mathbf{v}_j(\mathbf{z}^{k,j})\|.$$

If only $\tilde{\mathbf{x}}^{k,j}$ is accepted, then

$$\begin{aligned}\|\Delta_j^{k+1}\| &= \|\mathbf{x}_j^k + \tau \mathbf{v}_j(\mathbf{x}^{k,j}) + \sqrt{2\tau}\xi_j^k - \mathbf{z}_j^k\| \\ &\le \|\Delta_j^k\| + \tau M_v + \sqrt{2\tau}\|\xi_j^k\|\end{aligned}$$

Likewise, if only $\tilde{\mathbf{z}}_j^k$ is accepted, then

$$\|\Delta_j^{k+1}\| \le \|\Delta_j^k\| + \tau M_v + \sqrt{2\tau}\|\xi_j^k\|.$$

Summing over all four scenarios, we obtain

$$\begin{aligned}\mathbb{E}_{k,j}\|\Delta_j^{k+1}\| &\le \|\Delta_j^k\|\mathbb{P}_{k,j}(\text{reject both}) + \|\Delta_j^k + \tau \mathbf{v}_j(\mathbf{x}^{k,j}) - \tau \mathbf{v}_j(\mathbf{z}^{k,j})\|\mathbb{P}_{k,j}(\text{accept both}) \\ &\quad + \mathbb{E}_{k,j}(\|\Delta_j^k\| + \tau M_v + \sqrt{2\tau}\|\xi_j^k\|)\mathbf{1}_{\text{accept } \tilde{\mathbf{x}} \text{ or } \tilde{\mathbf{z}}} \\ &= \|\Delta_j^k + \tau \mathbf{v}_j(\mathbf{x}^{k,j}) - \tau \mathbf{v}_j(\mathbf{z}^{k,j})\| \\ &\quad + (\|\Delta_j^k\| - \|\Delta_j^k + \tau \mathbf{v}_j(\mathbf{x}^{k,j}) - \tau \mathbf{v}_j(\mathbf{z}^{k,j})\|)\mathbb{P}_{k,j}(\text{reject at least one}) \\ &\quad + \mathbb{E}_{k,j}(\tau M_v + \sqrt{2\tau}\|\xi_j^k\|)\mathbf{1}_{\text{accept } \tilde{\mathbf{x}} \text{ or } \tilde{\mathbf{z}}}.\end{aligned} \tag{1.9}$$

To prove the Proposition, it suffices to show that, for a constant M, the following two inequalities hold:

$$(\|\Delta_j^k\| - \|\Delta_j^k + \tau \mathbf{v}_j(\mathbf{x}^{k,j}) - \tau \mathbf{v}_j(\mathbf{z}^{k,j})\|)\mathbb{P}_{k,j}(\text{reject at least one}) \le \frac{1}{2} M\tau^{\frac{3}{2}} \sum_{i \in \mathcal{I}_j} \|\Delta_i^{k,j}\|, \tag{1.10}$$

$$\mathbb{E}_{k,j}(\tau M_v + \sqrt{2\tau}\|\xi_j^k\|)\mathbf{1}_{\text{accept } \tilde{\mathbf{x}} \text{ or } \tilde{\mathbf{z}}} \le \frac{1}{2} M\tau^{\frac{3}{2}} \sum_{i \in \mathcal{I}_j} \|\Delta_i^{k,j}\|. \tag{1.11}$$

The reason is that, if the above inequalities hold, we can use in (1.10) and (1.11) in (1.9), to obtain (1.8).

We will first show (1.11). Note that, by Lemma A.2 claim(2), there is a constant M_1 such that

$$\mathbb{P}_{k,j}(\text{accept } \tilde{\mathbf{x}} \text{ or } \tilde{\mathbf{z}}) \le M_1\tau \sqrt{\sum_{i \in \mathcal{I}_j} \|\mathbf{x}_i^{k,j} - \mathbf{z}_i^{k,j}\|^2} \le M_1\tau \sum_{i \in \mathcal{I}_j} \|\mathbf{x}_i^{k,j} - \mathbf{z}_i^{k,j}\| = M_1\tau \sum_{i \in \mathcal{I}_j} \|\Delta_i^{k,j}\|. \tag{1.12}$$

By Lemma A.2 claim(3)

$$\mathbb{E}_{k,j}\|\xi_j^k\|\mathbf{1}_{\text{accept } \tilde{\mathbf{x}} \text{ or } \tilde{\mathbf{z}}} \leq M_1\tau \sum_{i\in\mathcal{I}_j} \|\mathbf{x}_i^{k,j} - \mathbf{z}_i^{k,j}\| = M_1\tau \sum_{i\in\mathcal{I}_j} \|\Delta_i^{k,j}\|. \tag{1.13}$$

Since we assume that $\tau \leq 1$, we can plug (1.12) and (1.13) into the left hand side of (1.11), and find a constant M_2 so that

$$\begin{aligned}\mathbb{E}_{k,j}(\tau M_v + \sqrt{2\tau}\|\xi_j^k\|)\mathbf{1}_{\text{accept } \tilde{\mathbf{x}} \text{ or } \tilde{\mathbf{z}}} &= M_v\tau\mathbb{P}(\text{accept } \tilde{\mathbf{x}} \text{ or } \tilde{\mathbf{z}}) + \sqrt{2\tau}\mathbb{E}_{k,j}\|\xi_j^k\|\mathbf{1}_{\text{accept } \tilde{\mathbf{x}} \text{ or } \tilde{\mathbf{z}}} \\ &\leq \tau^{\frac{3}{2}} M_2 \sum_{i\in\mathcal{I}_j} \|\mathbf{x}_i^{k,j} - \mathbf{z}_i^{k,j}\| = \tau^{\frac{3}{2}} M_2 \sum_{i\in\mathcal{I}_j} \|\Delta_i^{k,j}\|.\end{aligned}$$

This proves (1.11).

We now use (1.11) to prove (1.10). By the triangular inequality and Assumptions 3.1 and 3.2, we have that

$$\left| \|\Delta_j^k + \tau\mathbf{v}_j(\mathbf{x}^{k,j}) - \tau\mathbf{v}_j(\mathbf{z}^{k,j})\| - \|\Delta_j^k\| \right| \leq \tau\|\mathbf{v}_j(\mathbf{x}^{k,j}) - \mathbf{v}_j(\mathbf{z}^{k,j})\| \leq \tau H_v \sum_{i\in\mathcal{I}_j} \|\Delta_i^{k,j}\|. \tag{1.14}$$

Moreover, by Lemma A.2 (1), there is a constant M_3 such that

$$\mathbb{P}_{k,j}(\text{reject at least one}) \leq M_3\sqrt{\tau}. \tag{1.15}$$

The product of (1.14) and (1.15) leads to (1.10):

$$(\|\Delta_j^k\| - \|\Delta_j^k + \tau\mathbf{v}_j(\mathbf{x}^{k,j}) - \tau\mathbf{v}_j(\mathbf{z}^{k,j})\|)\mathbb{P}_{k,j}(\text{reject at least one}) \leq H_v M_3 \tau^{\frac{3}{2}} \sum_{i\in\mathcal{I}_j} \|\Delta_i^{k,j}\|.$$

This concludes our proof. □

A.4 Contraction with block-wise log-concavity (Proof of Theorem 3.6)

We complete the proof of Theorem 3.6 by showing that the assumption of block-log-concavity implies that the coupled pair of MALA-within-Gibbs defines a contraction.

Lemma A.4. *Under Assumptions 3.1 and 3.2, for all coupled pairs* $\mathbf{x}, \mathbf{z}$*, and independently of the iteration number* k *(which we drop for convenience):*

$$\|\nabla^2_{\mathbf{x}_i,\mathbf{x}_j} \log\pi(\mathbf{x}) - \nabla^2_{\mathbf{z}_i,\mathbf{z}_j} \log\pi(\mathbf{z})\| = \|\nabla_{\mathbf{x}_i}\mathbf{v}_j(\mathbf{x}) - \nabla_{\mathbf{z}_i}\mathbf{v}_j(\mathbf{z})\| \leq H_v \sum_{l\in\mathcal{I}_{i,j}} \|\mathbf{x}_l - \mathbf{z}_l\|,$$

where $\mathcal{I}_{i,j} = \mathcal{I}_i \cap \mathcal{I}_j$ *has cardinality* $|\mathcal{I}_{i,j}| \leq S$*.*

Proof. By Assumption 3.1, we know that $\mathbf{v}_j(\mathbf{x})$ has no dependence on $\mathbf{x}_i$ if $i \notin \mathcal{I}_j$, so that we can write $\mathbf{v}_j(\mathbf{x}_{\mathcal{I}_j})$. Similarly, $\nabla_{\mathbf{x}_i}\mathbf{v}_j(\mathbf{x})$ can be written as $\nabla_{\mathbf{x}_i}\mathbf{v}_j(\mathbf{x}_{\mathcal{I}_j})$. For any $\mathbf{x}$ and $\mathbf{z}$, pick $\mathbf{y}, \mathbf{u} \in \mathbb{R}^d$ so that $\mathbf{y}_{\mathcal{I}_j} = \mathbf{x}_{\mathcal{I}_j}$, $\mathbf{y}_{\mathcal{I}_j^c} = \mathbf{z}_{\mathcal{I}_j^c}$ and $\mathbf{u}_{\mathcal{I}_i} = \mathbf{y}_{\mathcal{I}_i}$, $\mathbf{u}_{\mathcal{I}_i^c} = \mathbf{z}_{\mathcal{I}_i^c}$. Note that $\mathbf{u}$ will differ from $\mathbf{z}$ only at the blocks with indices in $\mathcal{I}_{i,j}$. Then, since $\nabla_{\mathbf{x}_i}\mathbf{v}_j(\mathbf{x}) = [\nabla_{\mathbf{x}_j}\mathbf{v}_i(\mathbf{x})]^T$ and by Assumption 3.2,

$$\begin{aligned}\|\nabla_{\mathbf{x}_i}\mathbf{v}_j(\mathbf{x}) - \nabla_{\mathbf{z}_i}\mathbf{v}_j(\mathbf{z})\| &= \|\nabla_{\mathbf{y}_i}\mathbf{v}_j(\mathbf{y}) - \nabla_{\mathbf{z}_i}\mathbf{v}_j(\mathbf{z})\| \\ &= \|\nabla_{\mathbf{y}_j}\mathbf{v}_i(\mathbf{y}) - \nabla_{\mathbf{z}_j}\mathbf{v}_i(\mathbf{z})\| \\ &= \|\nabla_{\mathbf{u}_j}\mathbf{v}_i(\mathbf{u}) - \nabla_{\mathbf{z}_j}\mathbf{v}_i(\mathbf{z})\| \\ &\leq H_v\|\mathbf{u} - \mathbf{z}\| \leq H_v \sum_{l\in\mathcal{I}_{i,j}} \|\mathbf{u}_l - \mathbf{z}_l\| = H_v \sum_{l\in\mathcal{I}_{i,j}} \|\mathbf{x}_l - \mathbf{z}_l\|.\end{aligned}$$

□

Proof of Theorem 3.6. From Proposition A.3, and using the fact that the differences of the pair of coupled MALA-within-Gibbs iterates in their blocks at the j-th block update is given by (1.7), we have a.s.,

$$\mathbb{E}_k\|\Delta_j^{k+1}\| \leq \mathbb{E}_k\|\Delta_j^k + \tau\mathbf{v}_j(\mathbf{x}^{k,j}) - \tau\mathbf{v}_j(\mathbf{z}^{k,j})\| + M\tau^{\frac{3}{2}} \sum_{i<j, i\in\mathcal{I}_j} \mathbb{E}_k\|\Delta_i^{k+1}\| + M\tau^{\frac{3}{2}} \sum_{i\geq j, i\in\mathcal{I}_j} \|\Delta_i^k\|. \quad (1.16)$$

Here, recall that $\mathbb{E}_k$ is the conditional expectation with respect to the information available before the k-th Gibbs cycle. To simplify notation, we define, for an arbitrary function g, $g[\mathbf{x},\mathbf{z}] = \int_0^1 g(s\mathbf{x}+(1-s)\mathbf{z})ds$. Recall that $\Delta^{k,j} = \mathbf{x}^{k,j} - \mathbf{z}^{k,j} = [\Delta_1^{k+1}, \Delta_2^{k+1}, \cdots, \Delta_{j-1}^{k+1}, \Delta_j^k, \cdots, \Delta_m^k]$. Also, we use

$$\nabla^2_{\mathbf{x}_j,\mathbf{x}} \log\pi(\mathbf{x}) = [\nabla^2_{\mathbf{x}_j,\mathbf{x}_1} \log\pi(\mathbf{x}), \ldots, \nabla^2_{\mathbf{x}_j,\mathbf{x}_m} \log\pi(\mathbf{x})],$$

to denote the j-th block-wise row of the Hessian log density.

First step: we decompose the first term of the right hand side of (1.16) into terms involving block-wise quantities. To this end, we first decompose

$$\begin{aligned}
\mathbf{v}_j(\mathbf{x}^{k,j}) - \mathbf{v}_j(\mathbf{z}^{k,j}) &= \nabla_{\mathbf{x}_j}\log\pi(\mathbf{x}^{k,j}) - \nabla_{\mathbf{z}_j}\log\pi(\mathbf{z}^{k,j}) \\
&= \int_0^1 \nabla^2_{\mathbf{x}_j,\mathbf{x}}\log\pi(\mathbf{x}^{k,j} + s\Delta^{k,j})\Delta^{k,j}ds \\
&= (\nabla^2_{\mathbf{x}_j,\mathbf{x}}\log\pi)[\mathbf{x}^{k,j},\mathbf{z}^{k,j}]\Delta^{k,j} \\
&= (\nabla^2_{\mathbf{x}_j,\mathbf{x}_j}\log\pi)[\mathbf{x}^{k,j},\mathbf{z}^{k,j}]\Delta_j^k + \sum_{i<j,i\in\mathcal{I}_j} (\nabla^2_{\mathbf{x}_j,\mathbf{x}_i}\log\pi)[\mathbf{x}^{k,j},\mathbf{z}^{k,j}]\Delta_i^{k+1} \\
&\quad + \sum_{i>j,i\in\mathcal{I}_J} (\nabla^2_{\mathbf{x}_j,\mathbf{x}_i}\log\pi)[\mathbf{x}^{k,j},\mathbf{z}^{k,j}]\Delta_i^k \qquad (1.17)\\
&= (\nabla^2_{\mathbf{x}_j,\mathbf{x}_j}\log\pi)[\mathbf{x}^{k,j},\mathbf{z}^{k,j}]\Delta_j^k + \sum_{i\neq j,i\in\mathcal{I}_j} (\nabla^2_{\mathbf{x}_j,\mathbf{x}_i}\log\pi)[\mathbf{x}^{k,j},\mathbf{z}^{k,j}]\Delta_i^k \\
&\quad + \sum_{i<j,i\in\mathcal{I}_J} (\nabla^2_{\mathbf{x}_j,\mathbf{x}_i}\log\pi)[\mathbf{x}^{k,j},\mathbf{z}^{k,j}](\Delta_i^{k+1} - \Delta_i^k).
\end{aligned}$$

Therefore, we can bound the first term on the right hand side of (1.16) by

$$\mathbb{E}_k\|\Delta_j^k + \tau\mathbf{v}_j(\mathbf{x}^{k,j}) - \tau\mathbf{v}_j(\mathbf{z}^{k,j})\| \leq \mathcal{D}_1 + \mathcal{D}_2 + \mathcal{D}_3, \quad (1.18)$$

where

$$\begin{aligned}
\mathcal{D}_1 &:= \mathbb{E}_k\left\|\Delta_j^k + \tau(\nabla^2_{\mathbf{x}_j,\mathbf{x}_j}\log\pi)[\mathbf{x}^{k,j},\mathbf{z}^{k,j}]\Delta_j^k\right\|, \\
\mathcal{D}_2 &:= \tau\sum_{i\neq j,i\in\mathcal{I}_j} \mathbb{E}_k\left\|(\nabla^2_{\mathbf{x}_j,\mathbf{x}_i}\log\pi)[\mathbf{x}^{k,j},\mathbf{z}^{k,j}]\right\| \|\Delta_i^k\|, \\
\mathcal{D}_3 &:= \tau\sum_{i<j,i\in\mathcal{I}_j} \mathbb{E}_k\left\|(\nabla^2_{\mathbf{x}_j,\mathbf{x}_i}\log\pi)[\mathbf{x}^{k,j},\mathbf{z}^{k,j}]\right\| \|\Delta_i^{k+1} - \Delta_i^k\|.
\end{aligned}$$

Second step: we replace $[\mathbf{x}^{k,j},\mathbf{z}^{k,j}]$ in $\mathcal{D}_1$ and $\mathcal{D}_2$ by $[\mathbf{x}^k,\mathbf{z}^k]$. The reason for doing this is that $\mathbf{x}^k$ and $\mathbf{z}^k$ are fixed for all j, which makes these quantities easier to analyze. Under the Lipschitz conditions for $\nabla_{\mathbf{x}_j,\mathbf{x}_i}\log\pi$ in Assumption 3.5, and recalling that

$$\mathbf{x}^{k,j} = [\mathbf{x}_1^{k+1}, \ldots, \mathbf{x}_{j-1}^{k+1}, \mathbf{x}_j^k, \ldots, \mathbf{x}_m^k],$$

we can bound $\|(\nabla^2_{\mathbf{x}_j,\mathbf{x}_i}\log\pi)[\mathbf{x}^{k,j},\mathbf{z}^{k,j}]-(\nabla^2_{\mathbf{x}_j,\mathbf{x}_i}\log\pi)[\mathbf{x}^k,\mathbf{z}^k]\|$ using Lemma A.4:

$$\begin{aligned}
&\|(\nabla^2_{\mathbf{x}_j,\mathbf{x}_i}\log\pi)[\mathbf{x}^{k,j},\mathbf{z}^{k,j}]-(\nabla^2_{\mathbf{x}_j,\mathbf{x}_i}\log\pi)[\mathbf{x}^k,\mathbf{z}^k]\|\\
&\leq\int_0^1\|\nabla^2_{\mathbf{x}_j,\mathbf{x}_i}\log\pi(s\mathbf{x}^{k,j}+(1-s)\mathbf{z}^{k,j})-\nabla^2_{\mathbf{x}_j,\mathbf{x}_i}\log\pi(s\mathbf{x}^k+(1-s)\mathbf{z}^k)\|ds\\
&\leq\int_0^1 H_v\left(\sum_{l\in\mathcal{I}_{j,i},l<j}s\|\mathbf{x}_l^{k+1}-\mathbf{x}_l^k\|+(1-s)\|\mathbf{z}_l^{k+1}-\mathbf{z}_l^k\|\right)ds\\
&\leq\frac{1}{2}H_v\sum_{l\in\mathcal{I}_{j,i},l<j}(\|\mathbf{x}_l^{k+1}-\mathbf{x}_l^k\|+\|\mathbf{z}_l^{k+1}-\mathbf{z}_l^k\|).
\end{aligned}\tag{1.19}$$

In (1.19), first note that the summation is over at most S terms, because the cardinality of the index set satisfies

$$|\{l:l\in\mathcal{I}_{j,i},l<j\}|\leq|\mathcal{I}_{j,i}|\leq S.$$

Second, note that either $\mathbf{x}_l^{k+1}=\mathbf{x}_l^k$ or $\mathbf{x}_l^{k+1}=\tilde{\mathbf{x}}_l^k=\mathbf{x}_l^k+\tau\mathbf{v}_l(\mathbf{x}^{k,l})+\sqrt{2\tau}\xi_l^k$, depending whether the proposal is rejected or not. Therefore,

$$\mathbb{E}_k\|\mathbf{x}_l^k-\mathbf{x}_l^{k+1}\|\leq\sqrt{\mathbb{E}_k\|\mathbf{x}_l^k-\tilde{\mathbf{x}}_l^k\|^2}\leq\sqrt{\tau^2\mathbb{E}_k\|\mathbf{v}_l(\mathbf{x}^{k,l})\|^2+2\tau q}\leq\tau M_v+\tau^{\frac{1}{2}}\sqrt{2q}.\tag{1.20}$$

The above bound also applies to $\mathbb{E}_k\|\mathbf{z}_l^k-\mathbf{z}_l^{k+1}\|$, for the same reasons. Therefore, by (1.19), for the constant $M_0:=H_vS(M_v\sqrt{\tau}+\sqrt{2q})$,

$$\mathbb{E}_k\tau\|(\nabla^2_{\mathbf{x}_j,\mathbf{x}_i}\log\pi)[\mathbf{x}^{k,j},\mathbf{z}^{k,j}]-(\nabla^2_{\mathbf{x}_j,\mathbf{x}_i}\log\pi)[\mathbf{x}^k,\mathbf{z}^k]\|\leq H_vS(M_v\tau^2+\sqrt{2q}\tau^{\frac{3}{2}})\leq M_0\tau^{\frac{3}{2}}.$$

This leads to following bound for $\mathcal{D}_1$ in (1.18),

$$\begin{aligned}
&\mathbb{E}_k\left\|\Delta_j^k+\tau(\nabla^2_{\mathbf{x}_j,\mathbf{x}_j}\log\pi)[\mathbf{x}^{k,j},\mathbf{z}^{k,j}]\Delta_j^k\right\|\\
&\leq\left\|\Delta_j^k+\tau(\nabla^2_{\mathbf{x}_j,\mathbf{x}_j}\log\pi)[\mathbf{x}^k,\mathbf{z}^k]\Delta_j^k\right\|+\tau\mathbb{E}_k\|(\nabla^2_{\mathbf{x}_j,\mathbf{x}_i}\log\pi)[\mathbf{x}^{k,j},\mathbf{z}^{k,j}]-(\nabla^2_{\mathbf{x}_j,\mathbf{x}_i}\log\pi)[\mathbf{x}^k,\mathbf{z}^k]\|\|\Delta_j^k\|\\
&\leq\left\|\Delta_j^k+\tau(\nabla^2_{\mathbf{x}_j,\mathbf{x}_j}\log\pi)[\mathbf{x}^k,\mathbf{z}^k]\Delta_j^k\right\|+M_0\tau^{\frac{3}{2}}\|\Delta_j^k\|\\
&\leq(1+\tau H_{j,j}[\mathbf{x}^k,\mathbf{z}^k]+M_0\tau^{\frac{3}{2}})\|\Delta_j^k\|.
\end{aligned}\tag{1.21}$$

To obtain the last inequality, we use Assumption 3.5 i).

Likewise, we can build an upper bound for $\mathcal{D}_2$ in (1.18).

$$\begin{aligned}
&\tau\mathbb{E}_k\left\|(\nabla^2_{\mathbf{x}_j,\mathbf{x}_i}\log\pi)[\mathbf{x}^{k,j},\mathbf{z}^{k,j}]\right\|\|\Delta_i^k\|\\
&\leq\tau\mathbb{E}_k\left\|(\nabla^2_{\mathbf{x}_j,\mathbf{x}_i}\log\pi)[\mathbf{x}^k,\mathbf{z}^k]\right\|\|\Delta_i^k\|+\tau\mathbb{E}_k\|(\nabla^2_{\mathbf{x}_j,\mathbf{x}_i}\log\pi)[\mathbf{x}^{k,j},\mathbf{z}^{k,j}]-(\nabla^2_{\mathbf{x}_j,\mathbf{x}_i}\log\pi)[\mathbf{x}^k,\mathbf{z}^k]\|\|\Delta_i^k\|\\
&\leq(\tau H_{j,i}[\mathbf{x}^k,\mathbf{z}^k]+M_0\tau^{\frac{3}{2}})\|\Delta_i^k\|.
\end{aligned}$$

In summary

$$\mathcal{D}_2=\sum_{i\neq j,i\in\mathcal{I}_j}\mathbb{E}_k\left\|(\nabla^2_{\mathbf{x}_j,\mathbf{x}_i}\log\pi)[\mathbf{x}^{k,j},\mathbf{z}^{k,j}]\right\|\|\Delta_i^k\|\leq S(\tau H_{j,i}[\mathbf{x}^k,\mathbf{z}^k]+M_0\tau^{\frac{3}{2}})\|\Delta_i^k\|.\tag{1.22}$$

Third step: we bound $\mathcal{D}_3$ in (1.18). Note that by Assumption 3.2, $\|\nabla^2_{\mathbf{x}_i,\mathbf{x}_j}\log\pi\|\leq H_v$, so that

$$\mathcal{D}_3\leq\tau H_v\sum_{i\in I_j}\mathbb{E}_k\|\Delta_i^{k+1}-\Delta_i^k\|=\tau H_v\sum_{i\in I_j}\mathbb{E}_k\|(\mathbf{x}_i^{k+1}-\mathbf{x}_i^k)-(\mathbf{z}_i^{k+1}-\mathbf{z}_i^k)\|.\tag{1.23}$$

Note that $\mathbf{x}_i^{k+1}$ is either $\mathbf{x}_i^k$ or $\tilde{\mathbf{x}}_i^k$ based on whether $\tilde{\mathbf{x}}_i^k$ is rejected or not. Thus,

$$\mathbf{x}_i^{k+1} - \mathbf{x}_i^k = \mathbb{1}_{\tilde{\mathbf{x}}_i^k \text{ is accepted}}(\tau \mathbf{v}_i(\mathbf{x}^{k,i}) + \sqrt{2\tau}\xi_i^k), \quad \mathbf{z}_i^{k+1} - \mathbf{z}_i^k = \mathbb{1}_{\tilde{\mathbf{z}}_i^k \text{ is accepted}}(\tau \mathbf{v}_i(\mathbf{z}^{k,i}) + \sqrt{2\tau}\xi_i^k).$$

This leads to

$$\mathbb{E}_k\|(\mathbf{x}_i^{k+1} - \mathbf{x}_i^k) - (\mathbf{z}_i^{k+1} - \mathbf{z}_i^k)\| = \mathcal{C}_1 + \mathcal{C}_2 + \mathcal{C}_3, \tag{1.24}$$

where

$$\begin{aligned}
\mathcal{C}_1 :&= \mathbb{E}_k \tau \|\mathbb{1}_{\tilde{\mathbf{x}}_i^k \text{ and } \tilde{\mathbf{z}}_i^k \text{are accepted}}(\mathbf{v}_i(\mathbf{x}^{k,i}) - \mathbf{v}_i(\mathbf{z}^{k,i}))\|,\\
\mathcal{C}_2 :&= \mathbb{E}_k \|\mathbb{1}_{\tilde{\mathbf{x}}_i^k \text{ is accepted } \tilde{\mathbf{z}}_i^k \text{is rejected}}(\tau \mathbf{v}_i(\mathbf{x}^{k,i}) + \sqrt{2\tau}\xi_i^k)\|,\\
\mathcal{C}_3 :&= \mathbb{E}_k \|\mathbb{1}_{\tilde{\mathbf{z}}_i^k \text{ is accepted } \tilde{\mathbf{x}}_i^k \text{is rejected}}(\tau \mathbf{v}_i(\mathbf{z}^{k,i}) + \sqrt{2\tau}\xi_i^k)\|.
\end{aligned}$$

To bound $\mathcal{C}_1$, we note that by (1.17) and Assumption 3.2, we have

$$\|\mathbf{v}_i(\mathbf{x}^{k,i}) - \mathbf{v}_i(\mathbf{z}^{k,i})\| \leq H_v \left(\sum_{l \geq i, l \in \mathcal{I}_i} \|\Delta_l^k\| + \sum_{l < i, l \in \mathcal{I}_i} \|\Delta_l^{k+1}\| \right).$$

Consequentially,

$$\mathcal{C}_1 \leq \mathbb{E}_k \tau \|\mathbf{v}_i(\mathbf{x}^{k,i}) - \mathbf{v}_i(\mathbf{z}^{k,i})\| \leq \tau H_v \left(\sum_{l \geq i, l \in \mathcal{I}_i} \|\Delta_l^k\| + \sum_{l < i, l \in \mathcal{I}_i} \mathbb{E}_k \|\Delta_l^{k+1}\| \right). \tag{1.25}$$

Next, we bound $\mathcal{C}_2$ using Lemma A.2, claims (2) and (3), so that for some constant M_0'

$$\begin{aligned}
\mathcal{C}_2 &\leq \tau \mathbb{E}_k \|\mathbb{1}_{\tilde{\mathbf{x}}_i^k \text{ is accepted } \tilde{\mathbf{z}}_i^k \text{is rejected}} \mathbf{v}_i(\mathbf{x}^{k,i})\| + \sqrt{2\tau}\mathbb{E}_k \|\mathbb{1}_{\tilde{\mathbf{x}}_i^k \text{ is accepted } \tilde{\mathbf{z}}_i^k \text{is rejected}} \xi_i^k\|\\
&\leq \tau M_v \mathbb{P}_k(\text{accept only one of } \tilde{\mathbf{x}}_i^k, \tilde{\mathbf{z}}_i^k) + \sqrt{2\tau}\mathbb{E}_k \mathbb{1}_{\text{accept only one of } \tilde{\mathbf{x}}_i^k, \tilde{\mathbf{z}}_i^k} \|\xi_i^k\|\\
&\leq M_0'(M_v\tau^2 + \sqrt{2\tau}\tau)\mathbb{E}_k \sqrt{\sum_{l \in \mathcal{I}_i} \|\mathbf{x}_l^{k,i} - \mathbf{z}_l^{k,i}\|^2}.
\end{aligned}$$

Also note that

$$\mathbb{E}_k \sqrt{\sum_{l \in \mathcal{I}_i} \|\mathbf{x}_l^{k,i} - \mathbf{z}_l^{k,i}\|^2} \leq \mathbb{E}_k \sum_{l \in \mathcal{I}_i} \|\mathbf{x}_l^{k,i} - \mathbf{z}_l^{k,i}\| = \sum_{l \geq i, l \in \mathcal{I}_i} \|\Delta_l^k\| + \sum_{l < i, l \in \mathcal{I}_i} \mathbb{E}_k \|\Delta_l^{k+1}\|.$$

Therefore,

$$\mathcal{C}_2 \leq M_0'(M_v\tau + \sqrt{2\tau}) \left(\sum_{l \geq i, l \in \mathcal{I}_i} \|\Delta_l^k\| + \sum_{l < i, l \in \mathcal{I}_i} \mathbb{E}_k \|\Delta_l^{k+1}\| \right). \tag{1.26}$$

The same bound holds for $\mathcal{C}_3$. Combining (1.23), (1.24), (1.25) and (1.26), we find that, for some constant M_1,

$$\mathcal{D}_3 \leq \sum_{i \in \mathcal{I}_j} \tau H_v(\mathcal{C}_1 + \mathcal{C}_2 + \mathcal{C}_3) \leq S M_1 \tau^{\frac{3}{2}} \left(\sum_{l \in \mathcal{I}_j^2} \|\Delta_l^k\| + \sum_{l \in \mathcal{I}_j^2} \mathbb{E}_k \|\Delta_l^{k+1}\| \right). \tag{1.27}$$

Here $\mathcal{I}_j^2 := \bigcup_{i \in \mathcal{I}_j} \mathcal{I}_i$; by the union bound of cardinality, $|\mathcal{I}_j^2| \leq S^2$. Also note that because $j \in \mathcal{I}_j$ so that $\mathcal{I}_j \subset \mathcal{I}_j^2$.

Summary of the first three steps: we bound the first term of the right hand side of (1.16) by bounding $\mathcal{D}_1$ with (1.21), $\mathcal{D}_2$ with (1.22), and $\mathcal{D}_3$ with (1.27) In conclusion, $\mathbb{E}_k\|\Delta_j^{k+1}\|$ can be bounded by

$$\mathbb{E}_k\|\Delta_j^{k+1}\| \leq (1+\tau H_{j,j}[\mathbf{x}^k,\mathbf{z}^k]+M_0\tau^{\frac{3}{2}})\|\Delta_j^k\| + \sum_{i\neq j, i\in\mathcal{I}_j}(\tau H_{j,i}[\mathbf{x}^k,\mathbf{z}^k]+M_0\tau^{\frac{3}{2}})\|\Delta_i^k\|$$
$$+\tau^{\frac{3}{2}}\left(\sum_{i\in\mathcal{I}_j^2} SM_1\mathbb{E}_k\|\Delta_i^k\| + \sum_{i\in\mathcal{I}_j^2} SM_1\mathbb{E}_k\|\Delta_i^{k+1}\| + M\sum_{i<j, i\in\mathcal{I}_j}\mathbb{E}_k\|\Delta_i^{k+1}\| + M\sum_{i\geq j, i\in\mathcal{I}_j}\|\Delta_i^k\|\right),$$

or, in a more succinct form,

$$\mathbb{E}_k\|\Delta_j^{k+1}\| \leq (1+\tau H_{j,j}[\mathbf{x}^k,\mathbf{z}^k]+M_0\tau^{\frac{3}{2}})\|\Delta_j^k\| + \sum_{i\neq j, i\in\mathcal{I}_j}(\tau H_{j,i}[\mathbf{x}^k,\mathbf{z}^k]+M_0\tau^{\frac{3}{2}})\|\Delta_i^k\|,$$
$$+\sum_{i\in\mathcal{I}_j^2}(SM_1+M)\tau^{\frac{3}{2}}\mathbb{E}_k\|\Delta_i^k\| + \sum_{i\in\mathcal{I}_j^2}(SM_1+M)\tau^{\frac{3}{2}}\mathbb{E}_k\|\Delta_i^{k+1}\|. \tag{1.28}$$

This provides us upper bounds on how $\|\Delta_j^k\|$ changes when the iteration number k increases.

Fourth step: we first rewrite (1.28) in a more compact matrix formulation. Define the following $\mathbb{R}^{m\times m}$ matrices by their entries

$$[\mathbf{M}_\tau]_{j,i} = \tau^{\frac{3}{2}}(SM_1+M)1_{i\in\mathcal{I}_j^2}, \quad [\mathbf{M}_\tau^0]_{j,i} = M_0\tau^{\frac{3}{2}}1_{i\in\mathcal{I}_j}.$$

Next, we define $D_k := [\|\Delta_1^k\|, \|\Delta_2^k\|, \cdots, \|\Delta_m^k\|]^T$. It can be verified that (1.28) is equivalent to the j-th row of the following vector inequality,

$$(\mathbf{I}-\mathbf{M}_\tau)\mathbb{E}_k D_{k+1} \preceq (\mathbf{I}+\mathbf{H}[\mathbf{x}^k,\mathbf{z}^k]\tau+\mathbf{M}_\tau+\mathbf{M}_\tau^0)D_k, \tag{1.29}$$

where we use the notation $\mathbf{x}\preceq\mathbf{y}$ to mean that the two vectors $\mathbf{x}$ and $\mathbf{y}$ satisfy $\mathbf{x}_i\leq\mathbf{y}_i$ for all blocks indexed by $i=1,\ldots,m$. Recall that for a matrix, its l_2-operator norm is bounded by its l_1-norm, see [26]. Thus,

$$\|\mathbf{M}_\tau\| \leq \tau^{\frac{3}{2}}\max_j|\mathcal{I}_j^2|SM_1 \leq \tau^{\frac{3}{2}}(M_1S+M)S^2, \quad \|\mathbf{M}_\tau^0\| \leq \tau^{\frac{3}{2}}\max_j|\mathcal{I}_j|M_0 \leq \tau^{\frac{3}{2}}M_0S. \tag{1.30}$$

For sufficiently small τ, $\|\mathbf{M}_\tau\|<1$, and we can thus write

$$(\mathbf{I}-\mathbf{M}_\tau)^{-1} = \mathbf{I}+\mathbf{M}_\tau+(\mathbf{M}_\tau)^2+(\mathbf{M}_\tau)^3+\cdots,$$

and

$$\|(\mathbf{I}-\mathbf{M}_\tau)^{-1}\| \leq \sum_{i=0}^{\infty}\|\mathbf{M}_\tau\|^i = \frac{1}{1-\tau^{\frac{3}{2}}(M_1S+M)S^2}. \tag{1.31}$$

Note that all entries of $\mathbf{M}_\tau$ are positive, therefore all entries of $(\mathbf{M}_\tau)^n$ and $(\mathbf{I}-\mathbf{M}_\tau)^{-1}$ are positive as well. Therefore, (1.29) leads to

$$\mathbb{E}_k D_{k+1} \preceq (\mathbf{I}-\mathbf{M}_\tau)^{-1}(\mathbf{I}+\mathbf{H}[\mathbf{x}^k,\mathbf{z}^k]\tau+\mathbf{M}_\tau+\mathbf{M}_\tau^0)D_k.$$

By law of total expectation, we have

$$\mathbb{E}D_k \preceq \mathbb{E}\prod_{l=0}^{k-1}(\mathbf{I}-\mathbf{M}_\tau)^{-1}(\mathbf{I}+\mathbf{H}[\mathbf{x}^l,\mathbf{z}^l]\tau+\mathbf{M}_\tau+\mathbf{M}_\tau^0)D_0.$$

Finally we provide an a.s. upper bound for $\|(\mathbf{I}-\mathbf{M}_\tau)^{-1}(\mathbf{I}+\mathbf{H}[\mathbf{x}^l,\mathbf{z}^l]\tau+\mathbf{M}_\tau+\mathbf{M}_\tau^0)\|$. By the log-concavity assumption, $\|\mathbf{I}+\mathbf{H}[\mathbf{x}^k,\mathbf{z}^k]\tau\|\leq 1-\tau\lambda_H$. Thus, by (1.30) and (1.31), for sufficiently small τ,

$$\|(\mathbf{I}-\mathbf{M}_\tau)^{-1}(\mathbf{I}+\mathbf{H}[\mathbf{x}^l,\mathbf{z}^l]\tau+\mathbf{M}_\tau+\mathbf{M}_\tau^0)\|\leq\frac{1-\tau\lambda_H+\tau^{\frac{3}{2}}(M_0S+M_1S^3+MS^2)}{1-\tau^{\frac{3}{2}}(M_1S+M)S^2}\leq(1-(1-\delta)\tau\lambda_H),\quad a.s..$$

Finally, we have that

$$\|\mathbb{E}D_k\|\preceq\left\|\mathbb{E}\prod_{k=0}^{k-1}(\mathbf{I}-\mathbf{M}_\tau)^{-1}(\mathbf{I}+\mathbf{H}[\mathbf{x}^l,\mathbf{z}^l]\tau+\mathbf{M}_\tau+\mathbf{M}_\tau^0)D_0\right\|\leq(1-(1-\delta)\tau\lambda_H)^k\|\mathbb{E}D_0\|.$$

□