Efficient MCMC Sampling with Dimension-Free Convergence Rate using ADMM-type Splitting

Maxime Vono  
IRIT/INP-ENSEEIHT  
University of Toulouse

Daniel Paulin  
Department of Statistics  
University of Oxford

Arnaud Doucet  
Department of Statistics  
University of Oxford

Abstract

Performing exact Bayesian inference for complex models is intractable. Markov chain Monte Carlo (MCMC) algorithms can provide reliable approximations of the posterior distribution but are computationally expensive for large datasets. A standard approach to mitigate this complexity consists in using subsampling techniques or distributing the data across a cluster. However, these approaches are typically unreliable in high-dimensional scenarios. We focus here on an alternative class of MCMC schemes exploiting a splitting strategy akin to the one used by the celebrated ADMM optimization algorithm. These methods, proposed recently in [45, 54], appear to provide empirically state-of-the-art performance. We generalize here these ideas and propose a detailed theoretical study of one of these algorithms known as the Split Gibbs Sampler. Under regularity conditions, we establish explicit dimension-free convergence rates for this scheme using Ricci curvature and coupling ideas. We demonstrate experimentally the excellent performance of these MCMC schemes on various applications.

1 Introduction

We are here interested in performing Bayesian inference for large datasets and potentially high-dimensional models. For complex models, the posterior distribution is intractable and needs to be approximated. Stochastic Variational Bayes approaches are popular in these scenarios as they are computationally rather cheap [29]. However, there is a lack of theoretical guarantees available for such approaches as the minimization problem one has to solve is typically not convex. Additionally, variational approximations tend to underestimate uncertainty. As a result, numerous MCMC schemes have been proposed over the past 5 years to perform Bayesian inference for large datasets; see [4] for a recent overview.

These methods can be loosely speaking divided into two groups: subsampling-based techniques and divide-and-conquer approaches. Subsampling-based approaches are MCMC techniques that only require accessing a subsample of the observations at each iteration: these include the popular Stochastic Gradient Langevin Dynamics (SGLD) [58, 20, 1, 9, 12], subsampling versions of the Metropolis–Hastings algorithm [3, 31, 4, 44, 16] and methods based on piecewise-deterministic MCMC schemes [7, 8]. However, all the subsampling methods accessing $O(1)$ data points at each iteration only provide reliable posterior approximations if they rely on some control variate ideas which require estimating the mode of the posterior and this posterior to be concentrated [20, 4, 1, 9, 12, 58, 16]. Practically, as pointed out in [4, 16], this means that such methods are of limited practical interest as they only work well in scenarios where the Bernstein-von Mises approximation of the target is excellent. Divide-and-conquer methods are methods which consider the common scenario where the data are distributed across a cluster. These schemes run independent MCMC chains to estimate “local” posteriors on each node of the cluster and then recombine these “local” posteriors to obtain an approximation of the full posterior [55, 39, 38, 56, 51, 50, 28]. However,
these methods often use parametric or kernel density approximations of the local posteriors so as to combine them. This can be unreliable in high-dimensional scenarios; see [4, 45] for a detailed discussion.

An alternative approach to perform MCMC, amenable to a distributed implementation, has been recently introduced in [45, 54]; see also [17, 2] for earlier related ideas. It is inspired by the well-known variable splitting technique used in optimization by the Alternating Direction Method of Multipliers (ADMM) [8]. In the sampling context, this corresponds to defining an instrumental hierarchical Bayesian model where the parameter of interest is artificially replicated as many times as one “splits” the target distribution. Experimentally, these methods appear promising but it is yet unclear how such MCMC schemes behave in high-dimensional scenarios.

Our contribution in this paper is three-fold. First, we provide a methodological extension of the splitting approach where it is not the original parameter of interest which is replicated but some potentially lower dimensional projections of it. Second, under standard assumptions such as Lipschitz gradient and convexity, we present non-asymptotic bounds on the total variation (TV) and 1-Wasserstein distances between the original posterior distribution and the distribution targeted by \( \theta \). We are interested in carrying out Bayesian inference about a parameter \( \theta \). An alternative approach to perform MCMC, amenable to a distributed implementation, has been potentially distributed over a cluster as in the logistic regression scenario, or the distribution can be represented via a potential function \( U \) for the likelihood \( \pi \), say for some matrices \( A_i \in \mathbb{R}^{d_i \times d} \) and potential functions \( U_i : \mathbb{R}^{d_i} \to \mathbb{R} \). We assume the potential \( U_i \) is independent on the subset \( y_i \) of the observations, potentially \( y_i = \{ \emptyset \} \). To simplify notation, this dependence is notationally omitted. We give two examples considered in Section 5.

**Example: Bayesian logistic regression.** We assign to \( \theta \) a Gaussian prior of precision \( 2 \tau I_d \). One possible factorization of the posterior considers \( b = n + 1 \) factors. We assign one potential per data by setting, for \( i \in [n] \), \( d_i = 1 \), \( A_i = x_i \), and \( U_i(u) = -y_i \log \left( h(u) \right) + \left( 1 - y_i \right) \log \left( 1 - h(u) \right) \), with \( h \) being the logistic link. The last potential corresponds to the prior, i.e. \( U_b(\theta) = \tau \| \theta \|^2 \) and \( A_b = I_d \).

**Example: Image inpainting.** The observed image is modeled as \( y = Hx + \epsilon \) where \( H \in \mathbb{R}^{m \times d} \) stands for a decimation matrix associated to a damaging binary mask and \( \epsilon \sim \mathcal{N}(0_m, \sigma^2 I_m) \). The original image \( \theta \) to recover is represented as a vector via lexicographic ordering. Under the total variation prior, a possible factorization of the posterior is to use a potential for the likelihood \( U_1(A_1 \theta) = \frac{(2 \sigma^2)^{-1}}{2} \| A_1 \theta - y \|^2 \) with \( A_1 = H \) and another one for the non-differential prior \( U_2(A_2 \theta) = \sum_{1 \leq i,j \leq \sqrt{d}} \| (A_2 \theta)_{i,j} \| \) where \( A_2 = \nabla \) is the 2D-discrete gradient, see [11] for a comprehensive overview of the total variation regularization.

Sampling from \( \pi \) in (1) is challenging because the number of data can be extremely large and potentially distributed over a cluster as in the logistic regression scenario, or the distribution can be non-differentiable as in the case of image inpainting. This rules out the application of standard techniques such as Hamiltonian Monte Carlo.
A simple sufficient assumption to ensure $\pi$ as in [45, 54], we introduce an instrumental Bayesian hierarchical model which will ease posterior computation. The idea is to introduce an auxiliary variable $z_i \in \mathbb{R}^d_i$ for each of the factor $i \in [b]$ such that, under an instrumental prior distribution, these variables are conditionally independent given $\theta$ and $z_i \sim \mathcal{N}(A_i \theta, \rho^2 I_{d_i})$ for some $\rho > 0$. We then consider the artificial posterior distribution $\pi_p(\theta, z_{1:b}) \propto \exp(-U(\theta, z_{1:b}))$ where

$$U(\theta, z_{1:b}) = \sum_{i=1}^b U_i(z_i) + \frac{||z_i - A_i \theta||^2}{2\rho^2}.$$ (2)

Figure 1 shows the directed acyclic graph (DAG) associated to this instrumental Bayesian hierarchical model. We could have considered an alternative prior for $z_i$ but this choice is motivated by the fact that the corresponding quadratic potential enjoys attractive properties (e.g., smoothness and strong convexity).

A simple sufficient assumption to ensure $\pi_p(\theta, z_{1:b})$ is a probability density function is that $U_i$ is bounded from below for every $j \in [b]$ and, for at least one $i \in [b]$, we have $d_i = d, A_i$ full rank, and $\exp(-U_i(z_i))$ integrable, see Appendix B.2 for more details. This assumption holds for the two examples considered in Section 5.

A key property of this artificial posterior distribution is that the resulting marginal posterior distribution $\pi_p(\theta) = \int \pi_p(\theta, z_{1:b})dz_{1:b}$ converges to the posterior distribution of interest $\pi(\theta)$ in total variation norm as $\rho \to 0$. This follows directly from the fact that $\mathcal{N}(z_i; A_i \theta, \rho^2 I_{d_i})$ converges towards the Dirac distribution $\delta_{A_i \theta}(z_i)$ when $\rho \to 0$ and Scheffé’s lemma [49].

This instrumental model is inspired by the splitting strategy exploited by ADMM, a popular distributed optimization technique [8]. Recall that ADMM addresses problems of the form $\min_\theta U_1(A_1 \theta) + U_2(A_2 \theta)$. It solves this problem in a distributed fashion by rewriting the objective as $\min_{\theta,z_1,z_2} U_1(z_1) + U_2(z_2)$ subject to $A_1 \theta = z_1, A_2 \theta = z_2$. The functions $U_1$ and $U_2$ can now be dealt with separately by considering the augmented Lagrangian $L_{\rho}(\theta, z_{1:2}, u_{1:2}) = \sum_{i=1}^2 U_i(z_i) + (2\rho^{-1}) ||A_i \theta - z_i + u_i||^2$, $u_i$ being a scaled dual variable, and alternating minimization and dual ascent steps.

The proposed instrumental potential (2) generalizes the approach in [45, 54]. In [45], only the case $A_i = I_{d_i}$ is considered so that $z_i \in \mathbb{R}^{d_i}$ where $d_i = d$. This can be very inefficient. In many applications, we can indeed define auxiliary variables $z_i$ living in $\mathbb{R}^{d_i}$ where $d_i \ll d$. In the logistic regression example presented in Section 5, we have $d_i = 1$ for $i \in [n]$ whereas $d = 785$ so simulation from $\pi_p(z_i; \theta)$ is much cheaper. Additionally, projecting the parameter of interest $\theta$ onto a lower-dimensional space can lead to conditionals for $z_i$ which are much easier to sample. This property is assessed experimentally in Sections 5.1 and 5.2.

2.3 MCMC algorithm

The main benefit of working with the artificial target distribution $\pi_p(\theta, z_{1:b})$ defined by (2) instead of $\pi(\theta)$ is the fact that, under $\pi_p$, the conditional distribution of the auxiliary variables $z_{1:b}$ given $\theta$...
Assumption 1

Algorithm 1: with

To prove these results, we shall introduce various regularity conditions in Assumption 1. This suggests using a Gibbs sampler to sample from the total variation (TV) and 1-Wasserstein distances between these distributions.

3.1 Results

In the previous section, we have established that the total variation between \( \pi_\rho(\theta) \) and \( \pi(\theta) \) goes to zero as \( \rho \to 0 \). We provide here more detailed results by establishing non-asymptotic bounds on both the total variation (TV) and 1-Wasserstein distances between these distributions.

3 Non-asymptotic results for the approximate model

In the previous section, we have established that the total variation between \( \pi_\rho(\theta) \) and \( \pi(\theta) \) goes to zero as \( \rho \to 0 \). We provide here more detailed results by establishing non-asymptotic bounds on both the total variation (TV) and 1-Wasserstein distances between these distributions.

3.1 Results

To prove these results, we shall introduce various regularity conditions in Assumption 1.

Assumption 1 (General assumptions).

\[
(A_0) \quad \inf_{z_j \in \mathbb{R}^d} U_j(z_j) > -\infty \text{ for every } j \in [b] \text{ (} U_j \text{ are bounded from below), and for at least one } i \in [b] \text{ we have } d_i = d, A_i \text{ is full rank, and } \exp(-U_i(z_i)) \text{ integrable on } \mathbb{R}^d.
\]

\[
(A_1) \quad U_i \text{ is } L_i\text{-Lipschitz, i.e., } \exists L_i > 0 \text{ such that } |U_i(z_i') - U_i(z_i)| \leq L_i \|z_i' - z_i\|, \forall z_i, z_i' \in \mathbb{R}^d.
\]

\[
(A_2) \quad U_i \text{ is continuously differentiable and admits a } M_i\text{-Lipschitz continuous gradient, i.e., } \exists M_i \geq 0 \text{ such that } \|\nabla U_i(z_i') - \nabla U_i(z_i)\| \leq M_i \|z_i' - z_i\|.
\]

\[
(A_3) \quad \int_{\mathbb{R}^d} \|\nabla U_i(A, \theta)\|^2 \pi(\theta) \, d\theta = M_i < +\infty.
\]

\[
(A_4) \quad U_i \text{ is convex, i.e. for every } \alpha \in [0, 1], z_i, z_i' \in \mathbb{R}^d, \text{ we have } U_i(\alpha z_i + (1-\alpha)z_i') \leq \alpha U_i(z_i) + (1-\alpha)U_i(z_i').
\]

\[
(A_5) \quad U_i \text{ is } m_i\text{-strongly convex, i.e., } \exists m_i > 0 \text{ such that } U_i(z_i) - \frac{m_i\|z_i\|^2}{2} \text{ is convex.}
\]

\[
(A_6) \quad d_1 = \ldots = d_b = d \text{ and } A_1 = \ldots = A_b = I_d.
\]
Table 1: Equivalent functions when \( \rho \to 0 \) of the non-asymptotic bounds given in Appendix C.

| Distance       | Assumptions | Equivalent of the upper bound |
|----------------|-------------|--------------------------------|
| \( \|\pi_{\rho} - \pi\|_{TV} \) | (A₁) \( \rho \sum_{i=1}^{b} 2\sqrt{d_i}L_i \) | |
|                 | (A₂), (A₃), (A₄) \( \rho^2 \sum_{i=1}^{d_i} M_i \) | |
| \( W_1(\pi_{\rho}, \pi) \) | (A₂), (A₅), (A₆) \( \min (\rho \sqrt{\bar{a}}, \frac{1}{2} \rho^2 \sqrt{M/d}) \) | |

Figure 2: From left to right: \( \|\pi - \pi_{\rho}\|_{TV} \), \( W_1(\pi, \pi_{\rho}) \), \( ||\nu P_{SGS}^t - \pi_{\rho}||_{TV} \) with \( \nu(\theta) = \mathcal{N}(\theta; \mu, \sigma^2/b) \) and \( W_1(\delta_{\theta_0} P_{SGS}^t, \pi_{\rho}) \) with \( \theta_0 = 0 \). The bounds shown in Sections 3.1 and 4.1 are also depicted.

Table 1 summarizes our non-asymptotic bounds on both TV and 1-Wasserstein distances. For sake of clarity and readability, only equivalent functions of the derived upper bounds on the TV distance, when \( \rho \) is sufficiently small, are given. Explicit formulas for these bounds and associated proofs are given in Appendix C. The results on the TV distance for \( b = 1 \) and \( d_1 = d \) are consistent with [40] who showed the same dependence w.r.t. \( \rho \) and \( d \) for a potential function and its smoothed version obtained by convolution with a Gaussian kernel.

3.2 Illustrations on a toy Gaussian model

The tightness of our bounds is illustrated on a toy Gaussian model for which a closed-form expression is available for \( \pi_{\rho} \). The target distribution is chosen as a scalar Gaussian \( \pi(\theta) = \mathcal{N}(\theta; \mu, \sigma^2/b) \) where \( b \geq 1 \) and \( \sigma > 0 \). In the sequel, we set \( \mu = 0, \sigma = 3 \) and \( b = 10 \). To satisfy the assumptions associated to each distance (see Table 1), we consider two splitting strategies.

Splitting strategy 1. Since the bound on \( \|\pi - \pi_{\rho}\|_{TV} \) is valid for any number of splitting operations, we set for all \( i \in [b], U_i(\theta) = (2\sigma^2)^{-1}(\theta - \mu)_+^2 \). The marginal of \( \theta \) under the instrumental hierarchical model in (2) has the closed-form expression \( \pi_{\rho}(\theta) = \mathcal{N}(\theta; \mu, (\sigma^2 + \rho^2)/b) \).

Splitting strategy 2. On the contrary, the bound in 1-Wasserstein distance has only been established for a single splitting. Hence, we set \( U(\theta) := U_1(\theta) = b(2\sigma^2)^{-1}(\theta - \mu)_+^2 \). Here, we have \( \pi_{\rho}(\theta) = \mathcal{N}(\theta; \mu, \sigma^2/b + \rho^2) \).

Figure 2 illustrates the bounds derived in Section 3.1 for both TV (with splitting strategy 1) and 1-Wasserstein (with splitting strategy 2) distances. Although being derived under general assumptions, these bounds appear to be quite tight, and especially the one associated to the 1-Wasserstein distance. Additional details are given in Appendix C.3.

4 Explicit convergence rates

4.1 Summary of our theoretical results

We are presenting here some convergence results for the SGS method. The precise statement of our results can be found in the appendix. We denote the Markov kernel of SGS in \( \theta \) by \( P_{SGS} \), i.e. \( P_{SGS}(\theta, \theta') = \int_{\mathbb{R}^b} \pi_{\rho}(z_{1:b} | \theta) \pi_{\rho}(\theta | z_{1:b}) dz_{1:b} \), where the conditionals are defined in (3) and (4).
This section illustrates the overall benefits of SGS for Bayesian inference. It shows that the proposed approach gives excellent performances and can be faster than state-of-the-art approaches. SGS is amenable to a distributed implementation but all the experiments have been run on a serial computer to emphasize that it is even beneficial in this context.

4.2 Illustrations on the toy Gaussian example

For any \( \theta \neq \theta' \in \mathbb{R}^d \), the coarse Ricci curvature of \( K(\theta, \theta') \) of \( \pi_{\text{SGS}} \) (defined in [41]) equals:

\[
K(\theta, \theta') = 1 - \frac{W_1(\pi_{\text{SGS}}(\theta', \cdot), \pi_{\text{SGS}}(\theta, \cdot))}{\|\theta - \theta'\|}.
\]

In Theorem 4, we show that, under Assumptions (A3) and (A6), this quantity is lower bounded by:

\[
K_{\text{SGS}} := \frac{\rho^2}{b} \sum_{i=1}^b \frac{m_i}{1 + m_i\rho^2}.
\]

(5)

An attractive property of this lower bound on the convergence rate \( K_{\text{SGS}} \) is that it is dimension-free, and only depends on \( b, \rho^2 \) and the strong convexity parameter \( m_i \). In Theorem 5, we show that the same curvature lower bound also holds for a stochastic version of SGS. This bound allows us to show that the absolute spectral gap of SGS is also lower bounded by \( K_{\text{SGS}} \) (see Corollary 1), and that for any initial distribution \( \nu \), we have:

\[
W_1(\nu P_{\text{SGS}}^t, \pi_p) \leq W_1(\nu, \pi_p) \cdot (1 - K_{\text{SGS}})^t,
\]

(6)

\[
\|\nu P_{\text{SGS}}^t - \pi_p\|_{\text{TV}} \leq \text{Var}_{\pi_p} \left( \frac{d\nu}{d\pi_p} \right) \cdot (1 - K_{\text{SGS}})^t.
\]

(7)

By combining our error bounds on \( W_1(\pi, \pi_p) \) for \( b = 1 \) (single splitting case discussed in Theorem 3) to the convergence bound (6), it is shown in Proposition 2 that in the strongly convex and smooth case, then we obtain a sample that has at most \( \epsilon \sqrt{d} / \sqrt{m} \) Wasserstein distance from \( \pi \) if we start from the minimizer \( \theta^* \) of \( U \), set \( \rho^2 = \max(\epsilon^2/(4m), \epsilon/\sqrt{mM}) \) and take \( t(\epsilon) \geq \log(3/\epsilon)/\log(1 + \max(\epsilon^2/4, \epsilon \sqrt{m/M})) \) iterations. In Corollary 2, we show that if \( \epsilon \leq d^{-1} \sqrt{m} / \sqrt{M} \), then sampling \( z_t \) given \( \theta \) can be performed by rejection sampling with \( O(1) \) expected evaluations of the function \( U \) and its gradient. As we can see on Table 2, this improves upon the available rates in the literature for sufficiently small precision \( \epsilon \). In Proposition 3, we state similar bounds in total variation distance, for the general \( b \geq 1 \) multiple splitting case.

4.2 Illustrations on the toy Gaussian example

Figure 2 illustrates the above convergence bounds on the toy Gaussian example introduced in Section 3.2. Technical details are given in Appendix D.5.

5 Experimental results

This section illustrates the overall benefits of SGS for Bayesian inference. It shows that the proposed approach gives excellent performances and can be faster than state-of-the-art approaches. SGS is amenable to a distributed implementation but all the experiments have been run on a serial computer to emphasize that it is even beneficial in this context.

### Table 2: Comparison of convergence rates in Wasserstein distance with the literature, starting from the minimizer \( \theta^* \) of the \( m \)-strongly convex and \( M \)-smooth potential \( U(\theta) \), with condition number \( \kappa = M/m \). SGS with single splitting is implemented based on rejection sampling. \( O^*() \) denotes \( O() \) up to polylogarithmic factors.

| Ref. | Method | Condition on \( \epsilon \) | Grad/func evals for \( W_1 \) err. |
|------|--------|-------------------|----------------------------------|
| [18] | Unadjusted Langevin | \( 0 < \epsilon \leq 1 \) | \( O^*(\kappa^2/\epsilon^2) \) |
| [15] | Underdamped Langevin | \( 0 < \epsilon \leq 1 \) | \( O^*(\kappa^2/\epsilon) \) |
| [19] | Underdamped Langevin | \( 0 < \epsilon \leq 1/\sqrt{\kappa} \) | \( O^*(\kappa^{3/2}/\epsilon) \) |
| [13] | Hamiltonian Dynamics | \( 0 < \epsilon \leq 1 \) | \( O^*(\kappa^{3/2}/\epsilon) \) |
| this paper | SGS with single splitting | \( 0 < \epsilon \leq 1/(d\sqrt{\kappa}) \) | \( O^*(\kappa^{1/2}/\epsilon) \) |
5.1 Bayesian regularized logistic regression

We consider the Bayesian logistic regression model detailed in Section 2.1 where we set $\tau = 5/2$. For $i \in [n]$, the $z_i$-conditionals are univariate and log-concave. We implemented both Rejection Sampling (see Corollary 2 in Section D.2) and an ULA-based approximate sampling scheme for sampling these conditionals. In Figure 3 in Section E.3 we show that these have the same convergence and classification performance. The ULA-based approach was somewhat faster in our Python implementation as it was easy to vectorize directly. We apply this model to five binary classification problems associated to the MNIST dataset with standard training (60000 examples) and test (10000 examples) sets. Figure 3 displays the results obtained by SGS (ULA-based implementation) and SGLD fixed point (SGLDFP) which stands for a control variate version of SGLD. On the first row, one observes that the convergence behavior of SGS w.r.t. $\rho$ is consistent with the results from Section 4 although $U_{i,i \in [n]}$ is not strongly convex. The larger $\rho$, the faster the convergence of the associated Markov chain towards its invariant distribution is. For example, setting $\rho = 1$ leads to similar classification performance as SGLDFP while convergence towards high-probability regions requires roughly 10 times less iterations. On the second row, we show experimentally that SGS converges roughly at the same speed (in CPU time) as SGLDFP towards high-probability regions. Table 3 gives the classification results associated to each MCMC algorithm on 5 binary handwritten digits classification problems. The minimum mean square estimate (MMSE) has been used to compute the probability to belong to each class. In this experiment, the SGS not only provides state-of-the-art classification scores (w.r.t. SGLDFP) but also outperforms it in most of the problems. See Appendix E for additional details and experiments.
Figure 4: Image inpainting. (top, from left to right) Original image, observation (60% of missing pixels depicted in black), MMSE obtained with SGS ($\rho = 1$) and associated 90% credibility intervals for the illuminance channel. (bottom) Potential $U = -\log \pi$, ISNR (dB) and MSE w.r.t. iteration $t$.

Table 4: Image inpainting. Performances of the SGS ($\rho = 1$) compared to MYULA for three images of celebrities. The performance criteria have been computed using the MMSE and averaged over 10 independent observations. $10^4$ samples have been used for each algorithm and burn-in periods of 5000 and 12500 iterations have been considered for SGS and MYULA, respectively.

| Image     | ISNR (dB) | MSE   | time (s) |
|-----------|-----------|-------|----------|
|           | MYULA     | SGS $\rho = 1$ | MYULA     | SGS $\rho = 1$ | MYULA     | SGS $\rho = 1$ |
| A. Keys   | 26.0 ± 0.3 | 25.8 ± 0.2 | 44.1 ± 3.0 | 45.8 ± 2.3 | 815 | 615 |
| D. Beckham| 26.3 ± 0.5 | 26.0 ± 0.3 | 31.6 ± 3.5 | 33.4 ± 1.9 | similar | similar |
| A. Mauresmo| 23.7 ± 0.3 | 23.4 ± 0.3 | 22.8 ± 1.6 | 24.0 ± 1.5 | similar | similar |

5.2 Image inpainting with non-differentiable prior

We consider here the image inpainting problem detailed in Section 2.1. The observed image $y \in \mathbb{R}^{m}$ has been obtained by only measuring 40% of the pixels of the original image, see Figure 4. We use MCMC algorithms to estimate the posterior mean along with credibility intervals. To this purpose, we compare SGS with the state-of-the-art proximal Moreau-Yosida ULA denoted MYULA [24] which has been developed to sample from non-smooth log-concave target distributions. Thanks to the splitting $\nabla \theta = z_2$, sampling from the $z_2$-conditional [3] involving the non-differentiable prior $\|\cdot\|$ has been conducted exactly using a data augmentation scheme akin to the one used in the Bayesian Lasso [43] while sampling from [4] has been done efficiently in the Fourier domain. In addition to accelerate the convergence towards high-probability regions (see Figure 4), this projection avoids the use of iterative algorithms [10] to approximate the proximity operator of the total variation norm as in MYULA. Three $250 \times 250$ ($d = 62500$) images of celebrities from the Labeled Faces in the Wild (LFW) database [30] have been considered. Figure 4 illustrates the performance of SGS compared to MYULA for the Beckham image. The performances in terms of image restoration have been measured by computing the ISNR and the MSE whose definitions are given in Appendix A. Again, the behavior of the proposed algorithm w.r.t. $\rho$ is consistent with the convergence rates shown in Section 4, although the TV prior is not strongly convex. In addition, for a suitable value of $\rho$ (here $\rho = 1$), SGS appears to be faster both in terms of total run time and number of iterations than MYULA while achieving state-of-the-art performance, see Table 4. Additional technical details and performance results for the three images are given in Appendix F.
Acknowledgements

This material is based upon work supported in part by the U.S. Army Research Laboratory and the U.S. Army Research Office, by the U.K. Ministry of Defence (MoD), and by the U.K. Engineering and Physical Research Council (EPSRC) under grant number EP/R013616/1. The authors thank the GdR ISIS and Rémi Bardenet from Université de Lille for funding MV’s visit to Oxford through the internationalization grant “Effet tunnel”. Part of this work has been supported by the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANITI).

A Guide to notations

| Notation | Description |
|----------|-------------|
| \( \Gamma(\cdot) \) | The gamma function. |
| \( \dim(u) \) | Dimension of the vector \( u \). |
| \( D_{-d}(\cdot) \) | The parabolic cylinder special function defined for all \( d > 0 \) and \( z \in \mathbb{R} \) by \( D_{-d}(z) = \exp(-z^2/4)\Gamma(d)^{-1} \int_0^{+\infty} e^{-xz-x^2/2x^d} dx \). |
| \( \| \cdot \| \) | The \( L^2 \) norm. |
| \( \| \cdot \|_\infty \) | The supremum norm. |
| \( \| \mu - \nu \|_{TV} \) | The total variation norm between two probability measures \( \mu \) and \( \nu \) on \( \mathbb{R}^d \).
| \( W_1(\mu, \nu) \) | The 1-Wasserstein distance between two probability measures \( \mu \) and \( \nu \) on \( \mathbb{R}^d \).
| \( N(\cdot; \mu, \sigma^2) \) | Gaussian probability distribution with mean \( \mu \) and variance \( \sigma^2 \). |
| \( \Pr(A) \) | Probability of the event \( A \). |
| \( \text{erf}(\cdot) \) | Error function. |
| \( \text{ISNR (dB)} \) | Improvement in signal-to-noise ratio defined by \( \text{ISNR} = 10 \log_{10} \left( \frac{\| \theta - y \|^2}{\| \theta - \hat{\theta} \|^2} \right) \) where \( y \) is the observed image, \( \theta \) is the original image and \( \hat{\theta} \) is the estimated one. |
| \( \text{MSE} \) | Mean squared error defined by \( \text{MSE} = d^{-1} \| \theta - \hat{\theta} \|^2 \) where \( d = \dim(\theta) \). |
| \( \text{prox}_f^\lambda \) | Proximity operator associated to the convex function \( f \) and defined for any \( \lambda > 0 \) as \( \text{prox}_f^\lambda(u) = \arg \min_x \left\{ f(x) + \frac{\|x - u\|^2}{2\lambda} \right\} \). |
| \( [n] \) | \( \{1, \ldots, n\} \). |
| \( \ker(M) \) | Null space of the linear map \( M \). |

B Additional details and proofs for Section 2

B.1 An extended state space Langevin diffusion process to target (2)

We point out here another possible approach to sample from the joint distribution \( \pi_y \) in (2) based on overdamped Langevin dynamics. The associated stochastic differential equation (SDE) writes

\[
\begin{align*}
\text{d} & \begin{pmatrix}
\theta_{t+} \\
\text{z}_{1,t+}
\end{pmatrix} = \\
& \begin{pmatrix}
\rho^{-2} \sum_{i=1}^{b} A_i^T \left( A_i \theta - z_{i,t} \right) \\
\rho^{-2} \left( \text{z}_{1,t} - A_1 \theta \right) + \nabla U_1(\text{z}_{1,t})
\end{pmatrix} \text{d}t + \sqrt{2} \begin{pmatrix}
d\xi_{t} \\
d\xi_{1,t}
\end{pmatrix},
\end{align*}
\]

(8)
where \((\xi_j)_{t \geq 0}\) and \((\xi'_t)_{t \geq 0}\) are independent \(d\)-dimensional and \(d_t\)-dimensional Brownian motions, respectively. By introducing the process \((s_t)_{t \geq 0} = (\theta_t, z_{1,t}, \ldots, z_{b,t})_{t \geq 0}\), the SDE (8) writes
\[
\text{d}s_t = -\nabla V(s_t) + \sqrt{2} \text{d}\xi'_t,
\]
where
\[
V(s_t) = \sum_{i=1}^{b} U_i(z_{i,t}) + \frac{1}{2\rho^2} \left\| A_i \theta_t - z_{i,t} \right\|^2,
\]
and \((\xi'_t)_{t \geq 0}\) is a \((d + k)\)-dimensional Brownian motion, where \(k = \sum_{i=1}^{b} d_i\).

Similarly to Algorithm [1], the SDE (8) leads to a divide-to-conquer implementation since each auxiliary variable \(z_{i,t}\) can be sampled independently from the others given the current iterate \(\theta_t\). An interesting advantage of working with (8) is that, contrary to SGS, the update of \(s_t\) can be undertaken in a fully parallel manner instead of a sequential one.

The theoretical analysis of the SDE (8) is out of the scope of this paper but will be considered in future work.

### B.2 Integrability of \(\pi_{\rho}\) and ergodicity of the SGS

**Proposition 1** (Integrability of \(\pi_{\rho}\) and ergodicity of SGS). Under Assumption (A0), \(\pi_{\rho}\) is integrable and SGS is \(\pi_{\rho}\)-irreducible and aperiodic.

**Proof.** First notice that using the conditional independence of \(z_j\) given \(\theta\) for all \(j \in [b]\), we have
\[
\pi_{\rho}(\theta) \propto \int_{\mathbb{R}^{d_b}} \exp(-U(\theta, z_{1:b})) \text{d}z_{1:b} = \prod_{j=1}^{b} \int_{\mathbb{R}^{d_i}} \exp \left(-U_j(z_j) - \frac{\| z_j - A_j \theta \|^2}{2\rho^2} \right) \cdot \frac{1}{(2\pi\rho^2)^{d_j/2}} \text{d}z_j.
\]
Now notice that for every \(j \in [b]\), we have
\[
\int_{\mathbb{R}^{d_i}} \exp \left(-U_j(z_j) - \frac{\| z_j - A_j \theta \|^2}{2\rho^2} \right) \cdot \frac{1}{(2\pi\rho^2)^{d_j/2}} \text{d}z_j \leq \exp \left(-\inf_{z_j} U_j(z_j) \right) \int_{\mathbb{R}^{d_i}} \exp \left(-\frac{\| z_j - A_j \theta \|^2}{2\rho^2} \right) \cdot \frac{1}{(2\pi\rho^2)^{d_j/2}} \text{d}z_j = \exp \left(-\inf_{z_j} U_j(z_j) \right).
\]
Using this bound for every \(j \neq i \) (as in Assumption (A0)), we have
\[
\prod_{j=1}^{b} \int_{\mathbb{R}^{d_i}} \exp \left(-U_j(z_j) - \frac{\| z_j - A_j \theta \|^2}{2\rho^2} \right) \cdot \frac{1}{(2\pi\rho^2)^{d_j/2}} \text{d}z_j 
\leq C \int_{\mathbb{R}^{d_i}} \exp \left(-U_i(z_i) - \frac{\| z_i - A_i \theta \|^2}{2\rho^2} \right) \cdot \frac{1}{(2\pi\rho^2)^{d_i/2}} \text{d}z_i,
\]
for some finite constant \(C\).

Integrating the latter term out in \(\theta\) gives
\[
\int_{\theta \in \mathbb{R}^d} \int_{\mathbb{R}^{d_i}} \exp \left(-U_i(z_i) - \frac{\| z_i - A_i \theta \|^2}{2\rho^2} \right) \cdot \frac{1}{(2\pi\rho^2)^{d_i/2}} \text{d}z_i \text{d}\theta
\]
The function $D$ is finite using the integrability condition on $z_i$. Hence $\pi_{\rho}(\theta)$ is integrable. The $\pi$-irreducibility and aperiodicity of the SGS follows because the SGS defined on the extended state space including $z_{1:b}$ is a Gibbs sampler with systematic scan, and it satisfies the positivity condition of Gibbs sampling (since the densities are always positive), see [47].

\[ \square \]

C Proofs for the results of Section 3

This section gives the proofs and technical details associated to the results presented in Section 3.

C.1 Non-asymptotic bounds for the total variation norm

This section is structured as follows. First, we prove the two non-asymptotic upper bounds on the total variation (TV) norm between $\pi_{\rho}$ and $\pi$ for $b = 1$, see [2]. Second, we give the proof for the equivalents listed in Table 1 for $b = 1$. Finally, we extend these results for an arbitrary number of blocks $b$.

C.1.1 Non-asymptotic bounds for the single splitting case

In all this section, we will assume that $b = 1$ in (1) and (2). In this case, the target $\pi$ and the approximate marginal $\pi_{\rho}$ can be re-written, for any $\theta \in \mathbb{R}^d$, as

\[
\pi(\theta) = \frac{\exp(-U_1(A_1\theta))}{\int_{\theta \in \mathbb{R}^d} \exp(-U_1(A_1\theta))d\theta}, \tag{11}
\]

\[
\pi_{\rho}(\theta) = \frac{\int_{z_1 \in \mathbb{R}^{d_1}} \exp \left( -U_1(z_1) - \frac{1}{2\rho^2} \|z_1 - A_1\theta\|^2 \right) dz_1}{\int_{z_1 \in \mathbb{R}^{d_1}} \int_{\theta \in \mathbb{R}^d} \exp \left( -U_1(z_1) - \frac{1}{2\rho^2} \|z_1 - A_1\theta\|^2 \right) dz_1 d\theta}. \tag{12}
\]

We begin by stating and proving the following lemma when $U_1$ satisfies (A1) in Assumption 1.

**Lemma 1.** Let $\pi$ and $\pi_{\rho}$ in (11) and (12), respectively. Let $U_1$ satisfy (A1) in Assumption 2. In addition, assume that $A_1$ has full column rank. Then, for any $\rho > 0$, we have

\[
\|\pi_{\rho} - \pi\|_{TV} \leq 1 - \Delta_{d_1}(\rho), \tag{13}
\]

where

\[
\Delta_{d_1}(\rho) = \frac{D_{-d_1}(L_1\rho)}{D_{-d_1}(-L_1\rho)}. \tag{14}
\]

The function $D_{-d_1}$ is the parabolic cylinder special function defined in Appendix A.

For sake of completeness, we detail the proof of this lemma which can also be found in [53] Theorem 2 & Appendix C] for the case $A_1 = I_d$.

**Proof.**

\[
\|\pi_{\rho} - \pi\|_{TV} = \frac{1}{2} \int_{\theta \in \mathbb{R}^d} \|\pi(\theta) - \pi_{\rho}(\theta)\|d\theta
\]
We now use the triangle inequality in (15) which leads to

\[
\frac{1}{2} \int_{\theta \in \mathbb{R}^d} \pi(\theta) \left| 1 - \frac{\pi_p(\theta)}{\pi(\theta)} \right| \, d\theta
\]

\[
= \frac{1}{2} \int_{\theta \in \mathbb{R}^d} \pi(\theta) \left| 1 - K(\theta) \frac{C_\pi}{C_{\pi_p}} \right| \, d\theta
\]  

(15)

where

\[
C_\pi = \int_{\theta \in \mathbb{R}^d} \exp(-U_1(A_1 \theta)) \, d\theta
\]  

(16)

\[
C_{\pi_p} = \int_{\theta \in \mathbb{R}^d} \int_{z_1 \in \mathbb{R}^d} \exp \left(-U_1(z_1) - \frac{1}{2\rho^2} \|z_1 - A_1 \theta\|^2\right) \, dz_1 \, d\theta
\]  

(17)

are the normalizing constants associated to \(\pi\) and \(\pi_p\), respectively, and

\[
K(\theta) = \int_{z_1 \in \mathbb{R}^d} \exp \left(U_1(A_1 \theta) - U_1(z_1) - \frac{1}{2\rho^2} \|A_1 \theta - z_1\|^2\right) \, dz_1.
\]  

(18)

The following result will be useful for the rest of the proof:

\[
\int_{\theta \in \mathbb{R}^d} K(\theta) \pi(\theta) \, d\theta = \frac{C_{\pi_p}}{C_\pi}.
\]  

(19)

Since \(U_1\) satisfies (A1) in Assumption 1, we have

\[
K(\theta) \leq \int_{z_1 \in \mathbb{R}^d} \exp \left(L_1 \|A_1 \theta - z_1\| - \frac{1}{2\rho^2} \|A_1 \theta - z_1\|^2\right) \, dz_1
\]

(with \(v = z_1 - A_1 \theta\)) = \(\int_{v \in \mathbb{R}^d} \exp \left(L_1 \|v\| - \frac{1}{2\rho^2} \|v\|^2\right) \, dv
\]

(with \(t = \|v\|\)) = \(\frac{2\pi^{d_1/2}}{\Gamma \left(\frac{d_1}{2}\right)} \int_0^\infty t^{d_1-1} \exp \left(L_1 t - \frac{1}{2\rho^2} t^2\right) \, dt.
\]

(20)

This integral admits a closed-form expression \([24\text{ Formula 3.462 1.}]\) by introducing the parabolic cylinder special function \(D_{-d_1}\) defined in Appendix A. Then, it follows

\[
K(\theta) \leq B_1,
\]

(21)

where

\[
B_1 = \frac{2\pi^{d_1/2} \rho^{d_1} \Gamma(d_1)}{\Gamma \left(\frac{d_1}{2}\right)} \exp \left(\frac{L_1^2 \rho^2}{4}\right) D_{-d_1}(-L_1 \rho).
\]

(22)

By (19) and (21), we also have

\[
\frac{C_{\pi_p}}{C_\pi} \geq \frac{1}{B_1}.
\]

(23)

We now use the triangle inequality in (15) which leads to

\[
\|\pi_p - \pi\|_{TV} \leq \frac{1}{2} \int_{\theta \in \mathbb{R}^d} \frac{C_\pi}{C_{\pi_p}} K(\theta) - \frac{1}{B_1} K(\theta) \left| \pi(\theta) - \frac{1}{2} \int_{\theta \in \mathbb{R}^d} K(\theta) \pi(\theta) \, d\theta \right| \, d\theta
\]

\[
= \frac{1}{2} \int_{\theta \in \mathbb{R}^d} \left(\frac{C_\pi}{C_{\pi_p}} K(\theta) - \frac{1}{B_1} K(\theta)\right) \pi(\theta) \, d\theta + \frac{1}{2} \int_{\theta \in \mathbb{R}^d} \frac{1}{B_1} K(\theta) - 1 \pi(\theta) \, d\theta
\]

\[
= \int_{\theta \in \mathbb{R}^d} \left(1 - \frac{K(\theta)}{B_1}\right) \pi(\theta) \, d\theta.
\]

(24)
Using one more time \((A_1)\) in Assumption 1 we have for all \(\theta, z_1,\)

\[- (U_1(z_1) - U_1(A_1\theta)) \geq -|U_1(z_1) - U_1(A_1\theta)| \geq - L_1 \| A_1\theta - z_1 \| \]

so that

\[ K(\theta) \geq \int_{z_1 \in \mathbb{R}^d} \exp \left( - L_1 \| A_1\theta - z_1 \| - \frac{1}{2\rho^2} \| A_1\theta - z_1 \|^2 \right) dz_1. \]  

(25)

With the same changes of variables as above, it follows

\[ K(\theta) \geq B_2, \]  

(26)

where

\[ B_2 = \frac{2\pi^{d_1/2} \rho^{d_1} \Gamma(d_1) \exp \left( \frac{L_1^2 \rho^2}{4} \right)}{\Gamma \left( \frac{d_1}{2} \right)} D_{-d_1}(L_1\rho). \]  

(27)

Then we have \( 1 - \frac{K(\theta)}{B_1} \leq 1 - \frac{B_2}{B_1} \) which combined with (24) yields

\[ \| \pi_\rho - \pi \|_{TV} \leq 1 - \frac{D_{-d_1}(L_1\rho)}{D_{-d_1}(-L_1\rho)}. \]  

(28)

We now prove another bound on the TV distance when \( U_1 \) satisfies \((A_2), (A_3)\) and \((A_4)\) in Assumption 1, see Lemma 2.

**Lemma 2.** Let \( \pi \) and \( \pi_\rho \) be as defined in (11) and (12), respectively. Let \( U_1 \) satisfy \((A_2), (A_3), (A_4)\) in Assumption 1. In addition, assume that \( A_1 \) has full column rank. Then, for all \( \rho > 0 \), we have

\[ \| \pi_\rho - \pi \|_{TV} \leq 1 - \left( 1 + 2\rho^2 M_1 \right)^{-d_1/2} \left( 1 - \frac{\rho^4 M_1 M_2}{1 + 2\rho^2 M_1} \right). \]  

(29)

**Proof.** The beginning of the proof follows the same lines as in the proof of Lemma 1. Hence, we have from (15) that

\[ \| \pi_\rho - \pi \|_{TV} = \frac{1}{2} \int_{\theta \in \mathbb{R}^d} \pi(\theta) \left| \int_{\theta' \in \mathbb{R}^d} \pi_\rho(\theta') \frac{K(\theta)}{C_{\pi_\rho}} d\theta' \right| d\theta. \]  

(30)

We now use the convexity of \( U_1 \) given by \((A_4)\) in Assumption 1 to write for all \( \theta \in \mathbb{R}^d, z_1 \in \mathbb{R}^d, \)

\[ U_1(A_1\theta) - U_1(z_1) \leq \nabla U_1(A_1\theta)^T (A_1\theta - z_1). \]  

(31)

By using (31), it follows that

\[ K(\theta) \leq \int_{z_1 \in \mathbb{R}^d} \exp \left( \nabla U_1(A_1\theta)^T (A_1\theta - z_1) - \frac{1}{2\rho^2} \| A_1\theta - z_1 \|^2 \right) dz_1 \]

\[ = \exp \left( \frac{\rho^2}{2} \| \nabla U_1(A_1\theta) \|^2 \right) \int_{z_1 \in \mathbb{R}^d} \exp \left( - \frac{1}{2\rho^2} \| z_1 - A_1\theta - \rho^2 \nabla U_1(A_1\theta) \|^2 \right) dz_1 \]

\[ = \exp \left( \frac{\rho^2}{2} \| \nabla U_1(A_1\theta) \|^2 \right) \left( 2\pi \rho^2 \right)^{d_1/2} = B_1(\theta). \]  

(32)

By using again \((A_4)\) in Assumption 1 we also have for all \( \theta \in \mathbb{R}^d, z_1 \in \mathbb{R}^d, \)

\[ U_1(A_1\theta) - U_1(z_1) \geq \nabla U_1(z_1)^T (A_1\theta - z_1). \]  

(33)
Then, (35) leads to
\[
K(\theta) \geq \int_{z_1 \in \mathbb{R}^d} \exp \left( \nabla U_1(z_1)^T (A_1 \theta - z_1) - \frac{1}{2\rho^2} \|A_1 \theta - z_1\|^2 \right) \, dz_1
\]
\[
= \int_{z_1 \in \mathbb{R}^d} \exp \left( \nabla U_1(z_1)^T (A_1 \theta - z_1) - \frac{1}{2\rho^2} \|A_1 \theta - z_1\|^2 \right) \times \exp \left( - \left( \nabla U_1(z_1) - \nabla U_1(z_1) \right)^T (A_1 \theta - z_1) \right) \, dz_1. \tag{34}
\]

We now use (A2) in Assumption [1] which leads to
\[
K(\theta) \geq \int_{z_1 \in \mathbb{R}^d} \exp \left( \frac{\rho^2}{2(1 + 2\rho^2 M_1)} \|\nabla U_1(z_1)\|^2 \right) \left( \frac{2\pi \rho^2}{1 + 2\rho^2 M_1} \right)^{d_1/2} \, dz_1 := B_2(\theta). \tag{35}
\]

We now apply the triangle inequality in (30) which yields
\[
\|\pi - \pi\|_{TV} \leq \frac{1}{2} \int_{\theta \in \mathbb{R}^d} \left| \frac{\pi}{C_{\pi}} - \frac{1}{B_1(\theta)} \right| K(\theta) \, d\theta + \frac{1}{2} \int_{\theta \in \mathbb{R}^d} \left| \frac{K(\theta)}{B_1(\theta)} - 1 \right| \pi(\theta) \, d\theta
\]
\[
= \frac{1}{2} \int_{\theta \in \mathbb{R}^d} \left| \frac{C_\pi}{C_{\pi_\rho}} - \frac{1}{B_1(\theta)} \right| K(\theta) \pi(\theta) \, d\theta + \frac{1}{2} \int_{\theta \in \mathbb{R}^d} \left( 1 - \frac{K(\theta)}{B_1(\theta)} \right) \pi(\theta) \, d\theta. \tag{36}
\]

Since $A_1$ has full column rank, the absolute value in the first term of (36) can be removed by noting that
\[
\frac{C_\pi}{C_{\pi_\rho}} = \frac{\int_{\theta \in \mathbb{R}^d} \exp(-U_1(A_1 \theta)) \, d\theta}{\int_{\theta \in \mathbb{R}^d} \int_{z_1 \in \mathbb{R}^d} \exp \left( -U_1(z_1) - \frac{1}{2\rho^2} \|z_1 - A_1 \theta\|^2 \right) \, dz_1 \, d\theta}
\]
\[
= \frac{\det (A_1^T A_1)^{-1/2} \int_{z_1 \in \mathbb{R}^d} \exp(-U_1(z_1)) \, dz_1}{\det (A_1^T A_1)^{-1/2} \int_{z_1 \in \mathbb{R}^d} \exp(-U_1(z_1)) \int_{u_1 \in \mathbb{R}^d} \exp \left( - \frac{1}{2\rho^2} \|z_1 - u_1\|^2 \right) \, du_1 \, dz_1}
\]
\[
= \left( \frac{2\pi \rho^2}{2} \right)^{-d_1/2} \exp \left( \frac{\rho^2}{2} \|\nabla U_1(A_1 \theta)\|^2 \right)
\]
\[
= \frac{1}{B_1(\theta)}. \tag{37}
\]

Then (36) becomes
\[
\|\pi - \pi\|_{TV} \leq \int_{\theta \in \mathbb{R}^d} \left( 1 - \frac{K(\theta)}{B_1(\theta)} \right) \pi(\theta) \, d\theta
\]
\[
\leq \int_{\theta \in \mathbb{R}^d} \left( 1 - \frac{B_2(\theta)}{B_1(\theta)} \right) \pi(\theta) \, d\theta
\]
\[
= 1 - \left( 1 + 2\rho^2 M_1 \right)^{-d_1/2} \int_{\theta \in \mathbb{R}^d} \exp \left( \frac{\rho^2}{2(1 + 2\rho^2 M_1)} - \frac{\rho^2}{2} \|\nabla U_1(A_1 \theta)\|^2 \right) \, d\theta \tag{38}
\]
We now use the fact that \(-\exp(-u) \leq u - 1\) for all \(u \geq 0\) which yields
\[
\|\pi_{\rho} - \pi\|_{TV} \leq 1 + \left(1 + 2\rho^2 M_1\right)^{-d_1/2} \left(\int_{\theta \in \mathbb{R}^d} \left(\frac{\rho^4 M_1 \|\nabla U_1(A_1, \theta)\|^2}{1 + 2\rho^2 M_1} - 1\right) \pi(\theta) \, d\theta - 1\right).
\]

(with \((A_3)\)) = 1 - \left(1 + 2\rho^2 M_1\right)^{-d_1/2} \left(1 - \frac{\rho^4 M_1 M_1}{1 + 2\rho^2 M_1}\right). \tag{40}

\[\blacksquare\]

### C.1.2 Equivalents for the single splitting case

The equivalent of the bound shown in Lemma 1 can be obtained by combining the results of the two following lemmas.

**Lemma 3.** When \(\rho \to 0\), we have
\[
\Delta_{d_1}(\rho) = 1 - \frac{2\sqrt{2} \Gamma\left(\frac{d_1 + 1}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right)} L_1 \rho + o(\rho). \tag{41}
\]

**Proof.** The parabolic cylinder function \(D_{-d_1}\) when \(d_1 > 0\) has the following expression
\[
D_{-d_1}(z) = \frac{\exp(-z^2/4)}{\Gamma(d_1)} \int_{0}^{\infty} e^{-x^2/2} x^{d_1-1} (1 - xz + o(z)) \, dx. \tag{42}
\]

In the limiting case when \(z \to 0\), a first order Taylor expansion of \(e^{-xz}\) gives
\[
D_{-d_1}(z) = \frac{\exp(-z^2/4)}{\Gamma(d_1)} \int_{0}^{\infty} e^{-x^2/2} x^{d_1-1} (1 - xz + o(z)) \, dx
= \frac{\exp(-z^2/4)}{\Gamma(d_1)} \left(\int_{0}^{\infty} e^{-x^2/2} x^{d_1-1} \, dx - z \int_{0}^{\infty} e^{-x^2/2} x \, dx + o(z)\right)
= \frac{\exp(-z^2/4)}{\Gamma(d_1)} \left(\Gamma\left(\frac{d_1}{2}\right) 2^{d_1/2-1} - 2 \Gamma\left(\frac{d_1 + 1}{2}\right) 2^{d_1/2-1/2} + o(z)\right). \tag{43}
\]

recording that \(\int_{0}^{\infty} e^{-x^2/2} x \, dx = \Gamma((d_1 + 1)/2) 2^{d_1/2-1/2} \) \cite{27} Formula 3.383 11]. Using \(43\) for \(z = \pm \rho L_1\) yields
\[
\Delta_{d_1}(\rho) = \frac{D_{-d_1}(L_1 \rho)}{D_{-d_1}(-L_1 \rho)} \left(\Gamma\left(\frac{d_1}{2}\right) 2^{d_1/2-1} - \rho L_1 \Gamma\left(\frac{d_1 + 1}{2}\right) 2^{d_1/2-1/2} + o(\rho)\right)
= \frac{\exp(-\rho L_1^2/4)}{\Gamma(d_1)} \left(\Gamma\left(\frac{d_1}{2}\right) 2^{d_1/2-1} + \rho L_1 \Gamma\left(\frac{d_1 + 1}{2}\right) 2^{d_1/2-1/2} + o(\rho)\right)
= \frac{\exp(-\rho L_1^2/4)}{\Gamma(d_1)} \Gamma\left(\frac{d_1}{2}\right) 2^{d_1/2-1} - \rho L_1 \Gamma\left(\frac{d_1 + 1}{2}\right) 2^{d_1/2-1/2} + o(\rho)
= \frac{\exp(-\rho L_1^2/4)}{\Gamma(d_1)} \Gamma\left(\frac{d_1}{2}\right) 2^{d_1/2-1} \left(1 + \rho \frac{L_1 \Gamma\left(\frac{d_1 + 1}{2}\right) \sqrt{2}}{\Gamma\left(\frac{d_1}{2}\right)} + o(\rho)\right).
\[
\left( 1 - \rho \frac{L_1 \Gamma \left( \frac{d_1 + 1}{2} \right) \sqrt{2}}{\Gamma \left( \frac{d_1}{2} \right)} + o(\rho) \right) \left( 1 - \rho \frac{L_1 \Gamma \left( \frac{d_1 + 1}{2} \right) \sqrt{2}}{\Gamma \left( \frac{d_1}{2} \right)} + o(\rho) \right) \\
= 1 - \frac{2 \sqrt{2} \Gamma \left( \frac{d_1 + 1}{2} \right)}{\Gamma \left( \frac{d_1}{2} \right)} L_1 \rho + o(\rho). \tag{44}
\]

**Lemma 4.** When \( d \to +\infty \), we have

\[
\frac{2 \sqrt{2} \Gamma \left( \frac{d_1 + 1}{2} \right)}{\Gamma \left( \frac{d_1}{2} \right)} L_1 \rho \quad \overset{d_1 \to +\infty}{\sim} \quad 2L_1 \rho d_1^{1/2}. \tag{45}
\]

**Proof.** The gamma function \( \Gamma \) can be expressed for all \( z > 0 \) as

\[
\Gamma(z) = \int_0^{+\infty} x^{z-1} e^{-x} \, dx.
\]

When \( z \) is large, Stirling-like approximations give the following equivalent for \( \Gamma(z + 1/2) \) and \( \Gamma(z) \):

\[
\Gamma(z + 1/2) \quad \overset{z \to +\infty}{\sim} \quad \sqrt{2\pi} z^{z} e^{-z}
\]

\[
\Gamma(z) \quad \overset{z \to +\infty}{\sim} \quad \sqrt{2\pi} z^{z - 1/2} e^{-z}.
\]

So that when \( d_1 \) is large

\[
\frac{2 \sqrt{2} \Gamma \left( \frac{d_1 + 1}{2} \right)}{\Gamma \left( \frac{d_1}{2} \right)} L_1 \rho \quad \overset{d_1 \to +\infty}{\sim} \quad \frac{2 \sqrt{2} \sqrt{2\pi} (d_1/2)^{d_1/2} e^{-d_1/2}}{\sqrt{2\pi} (d_1/2)^{d_1/2 - 1/2} e^{-d_1/2}} L_1 \rho \\
\quad \quad \overset{d_1 \to +\infty}{\sim} \quad \frac{2 \sqrt{2}}{d_1 (d_1/2)^{1/2} L_1 \rho} \\
\quad \quad \overset{d_1 \to +\infty}{\sim} \quad 2L_1 \rho d_1^{1/2}. \tag{48}
\]

Combining Lemmas [3] and [4] we obtain the following result.

**Lemma 5.** When \( \rho \) is sufficiently small and \( d_1 \) is sufficiently large, it follows under \((A_1)\) that

\[
\left\| \pi_\rho - \pi \right\|_{TV} \leq 2d_1^{1/2} \rho L_1 + o(\rho). \tag{49}
\]

The bound obtained in Lemma [4] admits the following equivalent when \( \rho \) is sufficiently small.

**Lemma 6.** When \( \rho \) is sufficiently small, it follows under \((A_2), (A_3), (A_4)\) that

\[
\left\| \pi_\rho - \pi \right\|_{TV} \leq \rho^2 d_1 M_1 + o(\rho^2). \tag{50}
\]

**Proof.** The proof follows from a straightforward first-order Taylor expansion in \( (40) \). \( \square \)

### C.1.3 Non-asymptotic bounds & equivalents for the multiple splitting case

We are now ready to extend the results of Appendices [C.1.1] and [C.1.2] to the multiple splitting case induced by \( \pi, \) resp. \( \pi_\rho, \) defined in [1], resp. [4].

---

\[\text{16}\]
Theorem 1. Let $\pi$ and $\pi_\rho$ be as defined in (1) and (2). For all $i \in [b]$, let $U_i$ and satisfy $(A_1)$ in Assumption 7. In addition, assume that the matrix

$$G := \begin{pmatrix} A_1 \\ \vdots \\ A_b \end{pmatrix} \in \mathbb{R}^{k \times d}, \text{ where } k = \sum_{i=1}^{b} d_i,$$

(51)

has full column rank. Then, for any $\rho > 0$, we have

$$\|\pi_\rho - \pi\|_{TV} \leq 1 - \prod_{i=1}^{b} \Delta_d(i)(\rho),$$

(52)

where

$$\Delta_d(i)(\rho) = \frac{D_d(L_i \rho)}{D_d(-L_i \rho)}.$$  

(53)

The function $D_d$ is the parabolic cylinder special function defined in Appendix A.

The equivalent of the bound in (52) when $\rho$ is sufficiently small writes

$$\|\pi_\rho - \pi\|_{TV} \leq 2\rho \sum_{i=1}^{b} d_i^{1/2} L_i + o(\rho).$$

(54)

Proof. This theorem is a straightforward extension of the proof of Lemma 1 by denoting that (18) becomes in the multiple splitting case

$$K(\theta) = \prod_{i=1}^{b} \int_{Z_i \in \mathbb{R}^{d_i}} \exp \left( U_i(A_i \theta) - U_i(z_i) - \frac{1}{2\rho^2} \|A_i \theta - z_i\|^2 \right) dz_i = \prod_{i=1}^{b} \mathcal{K}_i(\theta).$$

(55)

Then, bounding each term in (55) and following the proof of Lemma 1 detailed in Appendix C.1.1 completes the proof.

The equivalent of the upper bound can be obtained by using Lemma 5.

Theorem 2. Let $\pi$ and $\pi_\rho$ in (1) and (2), respectively. For all $i \in [0, n]$, let $U_i$ and satisfy $(A_2), (A_3), (A_4)$ in Assumption 7. Assume that the matrix

$$G := \begin{pmatrix} A_1 \\ \vdots \\ A_b \end{pmatrix} \in \mathbb{R}^{k \times d}, \text{ where } k = \sum_{i=1}^{b} d_i,$$

(56)

has full column rank. Then, for any $\rho > 0$, we have

$$\|\pi_\rho - \pi\|_{TV} \leq 1 - \prod_{i=1}^{b} \left( 1 + 2\rho^2 M_i \right)^{-d_i/2} \left( 1 - \sum_{i=1}^{b} \frac{\rho^4 M_i M_i}{1 + 2\rho^2 M_i} \right)$$

(57)

$$\leq \rho^2 \sum_{i=1}^{b} (d_i M_i + \rho^2 M_i M_i).$$

(58)

Proof. As in the proof of Lemma 2 using (55), we have

$$K(\theta) = \prod_{i=1}^{b} \mathcal{K}_i(\theta)$$

(59)

$$\leq \prod_{i=1}^{b} \int_{Z_i \in \mathbb{R}^{d_i}} \exp \left( \nabla U_i(A_i \theta)^T(A_i \theta - z_i) - \frac{1}{2\rho^2} \|A_i \theta - z_i\|^2 \right) dz_i$$
\[
\begin{align*}
&= \prod_{i=1}^{b} \left( \exp \left( \frac{\rho^2}{2} \| \nabla U_i(A_i \theta) \|^2 \right) \left( 2\pi \rho^2 \right)^{d_i/2} \right) := B_1(\theta), \\
\text{and} \\
K(\theta) &= \prod_{i=1}^{b} K_i(\theta) \\
&\geq \prod_{i=1}^{b} \int_{z_i \in \mathbb{R}^{d_i}} \exp \left( \nabla U_i(A_i \theta)^T (A_i \theta - z_i) - \left( \frac{1 + 2\rho^2 M_i}{2\rho^2} \right) \| A_i \theta - z_i \|^2 \right) \, dz_i \\
&= \prod_{i=1}^{b} \left( \exp \left( \frac{\rho^2}{2(1 + 2\rho^2 M_i)} \left\| \nabla U_i(A_i \theta) \right\|^2 \right) \left( \frac{2\pi \rho^2}{1 + 2\rho^2 M_i} \right)^{d_i/2} \right) := B_2(\theta). \tag{62}
\end{align*}
\]

As in the Lemma 2, we have
\[
\| \pi_\rho - \pi \|_{TV} \leq \int_{\theta \in \mathbb{R}^d} \left( 1 - \frac{B_2(\theta)}{B_1(\theta)} \right) \pi(\theta) \, d\theta \\
= 1 - \prod_{i=1}^{b} \left( 1 + 2\rho^2 M_i \right)^{-d_i/2} \cdot \int_{\theta \in \mathbb{R}^d} \exp \left( -\sum_{i=1}^{b} \frac{\rho^4 M_i \left\| \nabla U_i(A_i \theta) \right\|^2}{1 + 2\rho^2 M_i} \right) \pi(\theta) \, d\theta. \tag{63}
\]

Using the fact that $-\exp(-u) \leq u - 1$ for all $u \geq 0$, we have
\[
\| \pi_\rho - \pi \|_{TV} \leq 1 + \prod_{i=1}^{b} \left( 1 + 2\rho^2 M_i \right)^{-d_i/2} \cdot \int_{\theta \in \mathbb{R}^d} \left( \sum_{i=1}^{b} \frac{\rho^4 M_i \left\| \nabla U_i(A_i \theta) \right\|^2}{1 + 2\rho^2 M_i} - 1 \right) \pi(\theta) \, d\theta \\
= 1 - \left( \prod_{i=1}^{b} \left( 1 + 2\rho^2 M_i \right)^{-d_i/2} \right) \cdot \left( 1 - \sum_{i=1}^{b} \frac{\rho^4 M_i M_i}{1 + 2\rho^2 M_i} \right). \tag{64}
\]

Finally, notice that with the notation
\[
A := 1 - \left( \prod_{i=1}^{b} \left( 1 + 2\rho^2 M_i \right)^{-d_i/2} \right),
\]
\[
B := \sum_{i=1}^{b} \frac{\rho^4 M_i M_i}{1 + 2\rho^2 M_i},
\]

(66) implies that $\| \pi_\rho - \pi \|_{TV} \leq 1 - (1 - A)(1 - B) = A + B - AB \leq A + B$. Moreover, we have
\[
\log(1 - A) = \sum_{i=1}^{b} \frac{-d_i}{2} \log(1 + 2\rho^2 M_i),
\]
\[
1 - A = \exp \left( -\sum_{i=1}^{b} \frac{d_i}{2} \log(1 + 2\rho^2 M_i) \right) \geq 1 - \sum_{i=1}^{b} \frac{d_i}{2} \log(1 + 2\rho^2 M_i),
\]
\[
A \leq \sum_{i=1}^{b} \frac{d_i}{2} \log(1 + 2\rho^2 M_i) \leq \rho^2 \sum_{i=1}^{b} d_i M_i,
\]
\[
B = \sum_{i=1}^{b} \frac{\rho^4 M_i M_i}{1 + 2\rho^2 M_i} \leq \rho^4 \sum_{i=1}^{b} M_i M_i.
\]
C.2 Non-asymptotic bounds for the 1-Wasserstein distance

The following results show a Wasserstein error rate bound in the single splitting case \((b = 1)\).

**Theorem 3.** Assuming that \((A_b)\) holds, \(b = 1\), and \(U_1\) is twice differentiable, and satisfies that

\[
U_1(\theta) \geq a_1 + a_2||\theta|| \quad \text{and} \quad ||\nabla U_1(\theta)|| \leq a_3 + a_4||\theta||^\alpha
\]

for some \(a_2 > 0, \alpha > 0, a_1, a_3, a_4 \in \mathbb{R}\). Then we have

\[
W_1(\pi, \pi_\rho) \leq \min\left(\rho\sqrt{d}, \frac{1}{2}\rho^2 \int_\theta ||\nabla U_1(\theta)||\pi(\theta)d\theta\right).
\]

If \(U_1\) satisfies Assumptions \((A_2)\) and \((A_5)\) (strong convexity and gradient Lipschitz properties), then \((67)\) holds, and we have

\[
W_1(\pi, \pi_\rho) \leq \min\left(\rho\sqrt{d}, \frac{1}{2}\rho^2 \sqrt{M_1d}\right).
\]

**Proof.** Assume without loss of generality that \(U_1(\theta)\) is normalised, i.e. \(\int_{\theta \in \mathbb{R}^d} \exp(-U_1(\theta))d\theta = 1\) (if it is not, we can fix it by adding the logarithm of the normalising constant). Then the distribution

\[
\pi_\rho(\theta) = \frac{1}{(2\pi\rho^2)^{d/2}} \int_{z \in \mathbb{R}^d} \exp\left(-U_1(z) - \frac{||\theta - z||^2}{2\rho^2}\right) dz
\]

is the convolution of \(\pi(\theta) = \exp(-U_1(\theta))\) and a \(d\)-dimensional Gaussian random variable with mean zero and covariance \(\rho^2\mathbf{1}_d\). In particular, it is clear that

\[
\int_{\theta \in \mathbb{R}^d} \pi_\rho(\theta)d\theta = \int_{\theta \in \mathbb{R}^d} \frac{1}{(2\pi\rho^2)^{d/2}} \int_{\theta \in \mathbb{R}^d} \exp\left(-U_1(z) - \frac{||\theta - z||^2}{2\rho^2}\right) dzd\theta = \int_{\theta \in \mathbb{R}^d} \exp(-U_1(\theta)) = 1.
\]

The first part of the bound follows from the fact that the expectation of the norm of this Gaussian random variable is bounded by \(\rho\sqrt{d}\) (since the expectation of the square of the norm is \(\rho^2d\), this follows by Jensen’s inequality).

In order to obtain the second part, we are going to use the dual formulation of the 1-Wasserstein distance (see e.g. Remark 6.5 of [32]),

\[
W_1(\pi, \pi_\rho) = \sup_{g \in L_1(\mathbb{R}^d):||\nabla g||_\infty \leq 1} \left|\int_{\theta \in \mathbb{R}^d} g(\theta)\pi(\theta)d\theta - \int_{\theta \in \mathbb{R}^d} g(\theta)\pi_\rho(\theta)d\theta\right|
\]

\[
= \sup_{g \in C^1(\mathbb{R}^d):||\nabla g||_\infty \leq 1} \left|\int_{\theta \in \mathbb{R}^d} g(\theta)\pi(\theta)d\theta - \int_{\theta \in \mathbb{R}^d} g(\theta)\pi_\rho(\theta)d\theta\right|,
\]

where the second equality follows from the fact that differentiable functions \(g\) with \(||\nabla g||_\infty \leq 1\) are dense among 1-Lipschitz functions on \(\mathbb{R}^d\).

The evolution of a density \(\pi_\rho\) as we increase the variance \(\rho^2\) is known to follow the heat equation (see Section 2.4 of [32]),

\[
\frac{d}{d(\rho^2)}\pi_\rho(\theta) = \frac{1}{2}\Delta\pi_\rho(\theta),
\]

where \(\Delta\pi_\rho(\theta) = \sum_{i=1}^d \frac{\partial^2}{\partial\theta_i^2}\pi_\rho(\theta)\) denotes the Laplacian of \(\pi_\rho\). By integration, we obtain that

\[
\sup_{g \in C^1(\mathbb{R}^d):||\nabla g||_\infty \leq 1} \frac{d}{d(\rho^2)} \int_{\theta \in \mathbb{R}^d} g(\theta)\pi_\rho(\theta)d\theta = \sup_{g \in C^1(\mathbb{R}^d):||\nabla g||_\infty \leq 1} \frac{1}{2} \int_{\theta \in \mathbb{R}^d} g(\theta)\Delta\pi_\rho(\theta)d\theta.
\]
Now if we define the functional
\[ \mathcal{F}(\mu) := \sup_{g \in C^1(\mathbb{R}^d), \|\nabla g\|_{\infty} \leq 1} \frac{1}{2} \int_{\theta \in \mathbb{R}^d} g(\theta) \Delta \mu(\theta) d\theta. \]

Then it is easy to see that this is convex (\( \mathcal{F}((\mu + \nu)/2) \leq \frac{\mathcal{F}(\mu) + \mathcal{F}(\nu)}{2} \)) and shift-invariant (if \( \nu(x) = \mu(x - a) \) some constant \( a \in \mathbb{R}^d \), then \( \mathcal{F}(\nu) = \mathcal{F}(\mu) \)). Therefore it follows by the argument on pages 1-2 of [5] (monotonicity property of the heat semigroup for convex functionals) that \( \mathcal{F}(\pi_\rho) \leq \mathcal{F}(\pi) \) for every \( \rho \geq 0 \).

Initially, we have
\[ \mathcal{F}(\pi) = \sup_{g \in C^1(\mathbb{R}^d), \|\nabla g\|_{\infty} \leq 1} \frac{1}{2} \sum_{i=1}^d \int_{\theta \in \mathbb{R}^d} g(\theta_i) \frac{\partial^2}{\partial \theta_i^2} \pi(\theta) d\theta \]

After separating \( \theta \) to \( \theta_i \in \mathbb{R} \) and \( \theta_{-i} \in \mathbb{R}^{d-1} \) (denoting the rest of the coordinates), we have
\[
\int_{\theta \in \mathbb{R}^d} g(\theta) \frac{\partial^2}{\partial \theta_i^2} \pi(\theta) d\theta = \int_{\theta_{-i} \in \mathbb{R}^{d-1}} \left[ \int_{\theta_i \in \mathbb{R}} g(\theta) \frac{\partial^2}{\partial \theta_i^2} \pi(\theta) d\theta \right] d\theta_{-i},
\]
and now integration by parts and using condition (67) and the Lipschitz continuity of \( g \) leads to
\[
\int_{\theta_i \in \mathbb{R}} g(\theta) \frac{\partial^2}{\partial \theta_i^2} \pi(\theta) d\theta_i = \left[ -g(\theta) \frac{\partial}{\partial \theta_i} U_1(\theta) \cdot \exp(-U_1(\theta)) \right]_{\theta_i = -\infty}^{\theta_i = \infty} + \int_{\theta_i \in \mathbb{R}} \frac{\partial}{\partial \theta_i} g(\theta) \frac{\partial}{\partial \theta_i} U_1(\theta) \exp(-U_1(\theta)) d\theta_i = \int_{\theta_i \in \mathbb{R}} \frac{\partial}{\partial \theta_i} g(\theta) \frac{\partial}{\partial \theta_i} U_1(\theta) \pi(\theta) d\theta_i.
\]

By summing up in \( i \), we obtain that
\[ \mathcal{F}(\pi) \leq \frac{1}{2} \sup_{g \in C^1(\mathbb{R}^d), \|\nabla g\|_{\infty} \leq 1} \sum_{i=1}^d \int_{\theta \in \mathbb{R}^d} \frac{\partial}{\partial \theta_i} U_1(\theta) \frac{\partial}{\partial \theta_i} g(\theta) \pi(\theta) d\theta \leq \frac{1}{2} \int_{\theta \in \mathbb{R}^d} \|\nabla U_1(\theta)\| \pi(\theta) d\theta.
\]

Using the monotonicity property of \( F(\pi_\rho) \), now the second bound of the theorem follows based on formula (70).

Now we are going to consider the strongly convex and smooth \( U_1 \) case. In such situations, it is straightforward to see that condition (67) holds. For the integral of the norm of the gradient, we have by Jensen’s inequality
\[ \int_{\theta \in \mathbb{R}^d} \|\nabla U_1(\theta)\| \pi(\theta) d\theta \leq \left( \int_{\theta \in \mathbb{R}^d} \|\nabla U_1(\theta)\|^2 \pi(\theta) d\theta \right)^{1/2}. \]

For some index \( 1 \leq i \leq d \), we have
\[ \int_{\theta \in \mathbb{R}^d} \left( \frac{\partial}{\partial \theta_i} U_1(\theta) \right)^2 \pi(\theta) d\theta = \int_{\theta_{-i} \in \mathbb{R}^{d-1}} \left[ \int_{\theta_i \in \mathbb{R}} \left( \frac{\partial}{\partial \theta_i} U_1(\theta) \right)^2 \exp(-U_1(\theta)) d\theta_i \right] d\theta_{-i}, \]
and using integration by parts, and the conditions of strong convexity and smoothness, we have
\[
\int_{\theta_i \in \mathbb{R}} \left( \frac{\partial}{\partial \theta_i} U_1(\theta) \right)^2 \exp \left( -U_1(\theta) \right) d\theta_i
\]
\[
= \left[ -\exp \left( -U_1(\theta) \right) \right]_{\theta_i = -\infty}^{\theta_i = \infty} + \int_{\theta_i \in \mathbb{R}} \exp \left( -U_1(\theta) \right) \frac{\partial^2}{\partial \theta_i^2} U_1(\theta) d\theta_i
\]
\[
\leq \int_{\theta_i \in \mathbb{R}} \exp \left( -U_1(\theta) \right) M_1 d\theta_i,
\]
and by integrating this out according to \( \theta_i \) and summing up in \( i \), we obtain that
\[
\int_{\theta \in \mathbb{R}} d\|\nabla U_1(\theta)\|_2 \pi(\theta) d\theta \leq M_1 d,
\]
so the last claim of the theorem follows.

C.3 Additional details for the toy Gaussian example

The target distribution under consideration is the 1-dimensional Gaussian distribution with density
\[
\pi(\theta) = \mathcal{N} \left( \theta; \mu, \sigma^2 \right) \propto \exp \left( -\frac{b}{2\sigma^2} (\theta - \mu)^2 \right).
\]

C.3.1 Splitting strategy 1

By considering the splitting strategy 1 defined in Section 3.1, we set for \( i \in [b] \)
\[
U_i(\theta) = \frac{1}{2\sigma^2} (\theta - \mu)^2,
\]
leading to
\[
\pi(\theta) \propto \exp \left( -\sum_{i=1}^{b} U_i(\theta) \right).
\]

Following the instrumental hierarchical model introduced in Section 2.2, we introduce \( b \) univariate auxiliary variables \( z_i \) such that the corresponding approximate joint distribution \( \pi_\rho \) writes
\[
\pi_\rho(\theta, z_{1:b}) \propto \exp \left( -\sum_{i=1}^{b} \left[ U_i(z_i) + \frac{1}{2\rho^2} (z_i - \theta)^2 \right] \right). \tag{71}
\]

By a straightforward marginalization of \( z_{1:b} \) in (71), the marginal of interest associated to \( \theta \) is the Gaussian distribution with density defined by
\[
\pi_\rho(\theta) = \mathcal{N} \left( \theta; \mu, \frac{\sigma^2 + \rho^2}{b} \right).
\]

In dimension one, the 1-Wasserstein distance between two probability distributions \( \nu_1 \) and \( \nu_2 \) defined on \( \mathbb{R} \) can be expressed as
\[
W_1(\nu_1, \nu_2) = \int_{\mathbb{R}} |F_1(u) - F_2(u)| du,
\]
where \( F_1 \) and \( F_2 \) are the cumulative distribution functions (c.d.f.) associated to \( \nu_1 \) and \( \nu_2 \), respectively. The c.d.f. associated to a Gaussian distribution \( \mathcal{N}(\mu, \sigma^2) \) is
\[
F(u) = \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{u - \mu}{\sqrt{2\sigma}} \right) \right].
\]

Hence, using these two formulas, the 1-Wasserstein distance \( W_1(\pi, \pi_\rho) \) has been computed via numerical integration. Similarly, the total variation distance \( \|\pi - \pi_\rho\|_{\text{TV}} \) has also been computed using numerical integration.
C.3.2 Splitting strategy 2

By considering the splitting strategy 2 defined in Section 3.1, we set

\[ U_1(\theta) = \frac{b}{2\sigma^2}(\theta - \mu)^2, \]

leading to

\[ \pi(\theta) \propto \exp\left( -U_1(\theta) \right). \]

Following the instrumental hierarchical model introduced in Section 2.2, we introduce one univariate auxiliary variable \( z_1 \) such that the corresponding approximate joint distribution \( \pi_\rho \) writes

\[ \pi_\rho(\theta, z_1) \propto \exp\left( -U_1(z_1) - \frac{1}{2\rho^2}(z_1 - \theta)^2 \right). \quad (72) \]

By a straightforward marginalization of \( z_1 \) in (72), the marginal of interest associated to \( \theta \) is the Gaussian distribution with density defined by

\[ \pi_\rho(\theta) = \mathcal{N}(\theta; \mu, \sigma^2 \rho^2 + \frac{b}{\rho^2}). \]

D Explicit statements and proofs of the results of Section 4

D.1 Lower bounds on the coarse Ricci curvature

Our main result in this section is the following bound.

**Theorem 4** (Lower bound on curvature of SGS). Suppose that Assumptions \((A_5)\) and \((A_6)\) hold. Let

\[ K_{\text{SGS}} := \frac{\rho^2}{b} \sum_{i=1}^b \frac{m_i}{1 + m_i \rho^2}. \quad (73) \]

Then for the SGS, \( K(\theta, \theta') \geq K_{\text{SGS}} \) for any \( \theta \neq \theta' \in \mathbb{R}^d \). This implies in particular that for any initial distribution \( \nu \) on \( \mathbb{R}^d \), we have

\[ W_1(\nu \mathbf{P}^t_{\text{SGS}}, \pi_\rho) \leq W_1(\nu, \pi_\rho) \cdot (1 - K_{\text{SGS}})^t. \quad (74) \]

**Corollary 1** (Lower bound on the spectral gap of SGS). The SGS is a reversible Markov chain. Under Assumptions \((A_5)\) and \((A_6)\), its absolute spectral gap is lower bounded by \( K_{\text{SGS}} \) (see (73)).

**Proof.** The reversibility follows by a standard argument for data augmentation schemes given in Lemma 3.1 of [36]. The lower bound on the absolute spectral gap follows by Proposition 30 of [41].

An interesting feature of our curvature lower bound is that it also applies to the following stochastic version of SGS that only updates a random batch from the variables \( z_{1:b} \) in each iteration, which can be much faster than updating all of the variables. Let \( B_1, \ldots, B_r \subseteq \mathbb{N} \) be \( r \) subsets of the index set \( [b] = \{1, \ldots, b\} \), and \( p_1, \ldots, p_r \) be probability distribution on the subsets (i.e. \( 0 \leq p_i \leq 1 \), and \( p_1 + \ldots + p_r = 1 \)). For every \( 1 \leq j \leq r, 1 \leq i \leq b \), let \( w_{B_j}^i \geq 0 \) be weights satisfying the following assumptions.

**Assumption 2** (Homogeneity assumptions for SSGS).

\((H_0)\) For every \( i \in [b], j \in [r] \), \( w_{B_j}^i = 0 \) if \( i \notin B_j \).

\((H_1)\) For every \( j \in [r] \), \( \sum_{i \in [b]} w_{B_j}^i = b \).

\((H_2)\) For every \( i \in [b] \), \( \sum_{j \in [r]} p_j w_{B_j}^i = 1 \).
Then for the SSGS, we have 
\[ \mathbf{K}_{SSGS}, \mathbf{K}_{SSGS}', \mathbf{K}_{SSGS}'' \] which are subsampled versions of the mean and covariance used in Algorithm 1.

Theorem 5 (Lower bound on curvature of SSGS). Suppose that Assumptions \((A_5), (A_6)\) and \((H_0)-(H_2)\) hold. Let
\[ K_{SSGS} := \frac{\rho^2}{b} \sum_{i=1}^{b} \frac{m_i}{1 + m_i \rho^2}. \] (78)
Then for the SSGS, we have \( K(\mathbf{\theta}, \mathbf{\theta}') \geq K_{SSGS} \) for any \( \mathbf{\theta} \neq \mathbf{\theta}' \in \mathbb{R}^d \). This implies in particular that SSGS has a unique stationary distribution which we denote by \( \pi^{SSGS}_\rho \), and that for any initial distribution \( \nu \) on \( \mathbb{R}^d \), we have
\[ W_1 \left( \nu \mathbf{P}^{\rho}_{SSGS}, \pi^{SSGS}_\rho \right) \leq W_1 (\nu, \pi^{SSGS}_\rho) (1 - K_{SSGS})^t. \] (79)

**Remark.** A particular advantage of the weighted resampling formalism introduced here is that when \( U_1 \) is the prior distribution, we can include it with weight \( w_{B_j}^1 = 1 \) for each set \( B_j \), which is beneficial for reducing the variance of these stochastic estimates when the effect of the prior \( U_1 \) is more significant than any of the other terms. In Section 4, we compare the performance of SSGS with SGS for logistic regression on the MNIST dataset.

The following three lemmas are going to be used for the proof of Theorems 4 and 5. The first one will allow us to bound the Wasserstein distance of two log-concave distributions based on the differences between their gradients. This is achieved by coupling processes evolving according to the Unadjusted Langevin Algorithm (ULA).

**Lemma 7.** Let \( \mu \) and \( \mu' \) be two distributions on \( \mathbb{R}^n \) that are absolutely continuous with respect to the Lebesgue measure, and whose negative log-likelihoods are continuously differentiable, strongly convex and smooth (gradient-Lipschitz). Denote the strong convexity constants \( m(\mu), m(\mu') \) and smoothness constants \( M(\mu) \) and \( M(\mu') \). Then the Wasserstein distance of these two distributions

---

**Algorithm 2: Stochastic Split Gibbs Sampler (SSGS)**

**Input:** Potentials \( U_i \) for \( i \in [b] \), sets \( B_1, \ldots, B_r \subset [b] \), probabilities \( p_1, \ldots, p_r \) and weights \( (w_{B_j}^i)_{i \in [b], j \in [r]} \) satisfying Assumptions \((H_0)-(H_2)\), penalty parameter \( \rho \), initialization \( \mathbf{\theta}^{[0]} \) and nb. of iterations \( T \).

for \( t \leftarrow 1 \) to \( T \) do
- Sample the index of a set \( j \in [r] \) according to the probabilities \( p_1, \ldots, p_r \).
- for \( i \in B_j \) do
  - \( z_{i}^{[t]} \sim \pi_{\rho}(z_{i}|\mathbf{\theta}^{[t-1]}) \) (see Equation (5)).
- \( \mathbf{\theta}^{[t]} \sim \tilde{\pi}_{\rho}(\mathbf{\theta}|z_{B_j}^{[t]}) \) (see Equation (75)).
end

end

where
\[ \tilde{\pi}_{\rho}(\mathbf{\theta}|z_1, \ldots, z_{I_B}) = \mathcal{N}(\mathbf{\theta}; \mathbf{\mu}_{\rho}(z_{B_j}), \Sigma_{\rho}^{B_j}), \] (75)
with
\[ \Sigma_{\rho}^{B_j} = \rho^2 \left( \sum_{i \in B_j} w_{B_j}^i A_i^T A_i \right)^{-1}, \] (76)

\[ \mathbf{\mu}_{\rho}(z_{B_j}) = \left( \sum_{i \in B_j} w_{B_j}^i A_i^T z_i \right) \sum_{i \in B_j} w_{B_j}^i A_i^T A_i, \] (77)

which are subsampled versions of the mean and covariance used in Algorithm 1.

Then our next result shows that the same convergence rates hold for SSGS as for SGS.

**Theorem 5 (Lower bound on curvature of SSGS).** Suppose that Assumptions \((A_5), (A_6)\) and \((H_0)-(H_2)\) hold. Let
\[ K_{SSGS} := \frac{\rho^2}{b} \sum_{i=1}^{b} \frac{m_i}{1 + m_i \rho^2}. \] (78)

Then for the SSGS, we have \( K(\mathbf{\theta}, \mathbf{\theta}') \geq K_{SSGS} \) for any \( \mathbf{\theta} \neq \mathbf{\theta}' \in \mathbb{R}^d \). This implies in particular that SSGS has a unique stationary distribution which we denote by \( \pi^{SSGS}_\rho \), and that for any initial distribution \( \nu \) on \( \mathbb{R}^d \), we have
\[ W_1 \left( \nu \mathbf{P}^{\rho}_{SSGS}, \pi^{SSGS}_\rho \right) \leq W_1 (\nu, \pi^{SSGS}_\rho) (1 - K_{SSGS})^t. \] (79)

**Remark.** A particular advantage of the weighted resampling formalism introduced here is that when \( U_1 \) is the prior distribution, we can include it with weight \( w_{B_j}^1 = 1 \) for each set \( B_j \), which is beneficial for reducing the variance of these stochastic estimates when the effect of the prior \( U_1 \) is more significant than any of the other terms. In Section 4, we compare the performance of SSGS with SGS for logistic regression on the MNIST dataset.
where the two steps share the same standard normal noise variable $\xi$. Then by Lemma 8 and Corollary 21 of [41], it follows that for $\delta \leq \frac{2}{M(\mu') + m(\mu)}$, we have

$$W_1(\mathcal{L}(Z_1'), \mu') \leq (1 - m(\mu') \delta) W_1(\mu', \mu').$$

Here $\mathcal{L}(Z_1')$ refers to the law of $Z_1'$. Moreover, by the coupling (using the fact that $Z_0 = Z_0'$) it follows that

$$W_1(\mathcal{L}(Z_1'), \mu') = W_1(\mathcal{L}(Z_1'), \mathcal{L}(Z_1)) \leq \mathbb{E}(\|\delta \nabla \log (\mu(Z_0)) - \delta \nabla \log (\mu'(Z_0))\|)$$

$$= \delta \int_{z \in \mathbb{R}^n} \|\nabla \log (\mu)(z) - \nabla \log (\mu')(z)\| \mu(z) dz.$$

By combining the above two results using triangle inequality, we obtain that

$$W_1(\mu', \mu') \leq (1 - m(\mu') \delta) W_1(\mu', \mu') + \delta \int_{z \in \mathbb{R}^n} \|\nabla \log (\mu)(z) - \nabla \log (\mu')(z)\| \mu(z) dz,$$

hence by rearrangement

$$W_1(\mu', \mu') \leq \frac{\int_{z \in \mathbb{R}^n} \|\nabla \log (\mu)(z) - \nabla \log (\mu')(z)\| \mu(z) dz}{m(\mu')}.$$

The result now follows by letting $\delta \to 0$, using the triangle inequality and noticing that $W_1(\mu', \mu) \to 0$ and $W_1(\mu', \mu') \to 0$ as $\delta \to 0$ (this follows from example from Theorem 1 of [18]).

**Lemma 8** (Lower bound on the coarse Ricci curvature for ULA). Let $\mu$ be a distribution on $\mathbb{R}^n$ with density proportional to $\exp(-U(z))$, such that $U$ is $m$-strongly convex and $M$-gradient Lipschitz for some constants $0 < m \leq M$. Let $P_{\delta}$ denote the Markov kernel of the Markov chain evolving according to ULA

$$Z_{k+1} = Z_k - \delta \nabla U(Z_k) + \sqrt{2\delta} \xi_k,$$

where $\xi_k$ are i.i.d. d-dimensional standard normal random variables. Let

$$K_{\delta} := 1 - \max(|1 - \delta m|, |1 - \delta M|)$$

$$= \delta m \quad \text{if} \quad \delta \leq \frac{2}{m + M},$$

then $P_{\delta}$ satisfies that $K(z, z') \geq K_{\delta}$ for every disjoint $z \neq z' \in \mathbb{R}^n$.

**Proof.** Since both $P_{\delta}(z, \cdot)$ and $P_{\delta}(z', \cdot)$ are Gaussians with covariance $2\delta I_d$, their Wasserstein distance equals the distance of their means, i.e.

$$W_1(P_{\delta}(z, \cdot), P_{\delta}(z', \cdot)) = \| z - z' - (\nabla U(z) - \nabla U(z')) \delta \|$$
by the mean value theorem, there is a \( w \in \mathbb{R}^n \) such that

\[
= \| (I_d - \delta \nabla^2 U(w))(z - z') \| \leq \| I_d - \delta \nabla^2 U(w) \| \| z - z' \|
\]

\[
\leq \max(1 - \delta m, 1 - \delta M) \cdot \| x - y \|
\]

where the last step follows using the fact that \( (1 - \delta M) I_n \leq I_n - \delta \nabla^2 U(w) \leq (1 - \delta m) I_n \), implying that all of the eigenvalues of \( I_n - \delta \nabla^2 U(w) \) are in the interval \([1 - \delta M, 1 - \delta m]\). The result now follows by the definition of the coarse Ricci curvature.

\( \square \)

**Lemma 9.** Let \( \theta, \theta' \in \mathbb{R}^d \) be two parameter values, and \( \mu_i, \) resp. \( \mu'_i, \) denotes the conditional distributions of \( z_i \) given \( \theta \) under \( \pi_\rho, \) resp. \( \theta' \). Then under Assumptions \((A_5)\) and \((A_6)\), we have

\[
W_1(\mu_i, \mu'_i) \leq \frac{1}{1 + \rho^2 m_i} \| \theta - \theta' \|.
\]

**Proof.** We have \( \mu_i(z) \propto \exp \left( -U_i(z) - \frac{\| \theta - z \|^2}{2\rho^2} \right) \) and \( \mu'_i(z) \propto \exp \left( -U_i(z) - \frac{\| \theta' - z \|^2}{2\rho^2} \right) \).

Lemma [7] requires the smoothness (gradient Lipschitz) property, so it cannot be applied directly to these potentials under our assumptions. To overcome this difficulty, we are going to use the Moreau-Yosida envelope of \( U_i \) (see e.g. [24]), defined for any \( \lambda > 0 \) as

\[
U_i^\lambda(z) := \min_{y \in \mathbb{R}^d} \left\{ U_i(y) + (2\lambda)^{-1} \| y - z \|^2 \right\},
\]

where \( \lambda > 0 \) is a regularisation parameter. By Theorem 1.25 of [48], it follows that \( U_i^\lambda \) converges pointwise to \( U_i \), i.e. for every \( z \in \mathbb{R}^d \),

\[
\lim_{\lambda \to 0} U_i^\lambda(z) = U_i(z).
\]

Moreover, from Proposition 12.19 of [48] and Theorem 2.2 of [35] it follows that \( U_i^\lambda \) is \( \lambda^{-1} \) gradient Lischitz and \( \frac{m_i}{1 + \lambda m_i} \)-strongly convex.

Let \( \mu_i^\lambda(z) \propto \exp \left( -U_i^\lambda(z) - \frac{\| \theta - z \|^2}{2\rho^2} \right) \) and \( \mu'_i^\lambda(z) \propto \exp \left( -U_i^\lambda(z) - \frac{\| \theta' - z \|^2}{2\rho^2} \right) \), then we have

\[
\| \nabla \log(\mu_i^\lambda(z)) - \nabla \log(\mu'_i^\lambda(z)) \| = \frac{\| \theta - \theta' \|}{\rho^2}.
\]

Since \( -\log \mu_i(z) \) and \( -\log \mu'_i(z) \) are \( \frac{m_i}{1 + \lambda m_i} + \frac{1}{\rho^2} \)-strongly convex and \( \frac{1}{\lambda} + \frac{1}{\rho^2} \)-smooth, it follows by Lemma [7] that we have

\[
W_1(\mu_i^\lambda, \mu'_i^\lambda) \leq \frac{\| \theta - \theta' \|}{1 + \rho^2 m_i/(1 + m_i \lambda)}.
\]

To complete the proof, we still need to bound \( W_1(\mu_i^\lambda, \mu_i) \). Note that by Theorem 6.15 of [52], we have

\[
W_1(\mu_i^\lambda, \mu_i) \leq \int_{z \in \mathbb{R}^d} \| z - \theta \| \| \mu_i(z) - \mu'_i(z) \| \mathrm{d}z.
\]

Note that \( |\mu_i(z) - \mu'_i(z)| \leq \mu_i(z) + \mu'_i(z) \). Moreover, from the definition of the Moreau-Yosida envelope \( U_i^\lambda \), it follows that \( \mu_i^\lambda(z) < U_i^\lambda(z) \) for \( \lambda' < \lambda \), hence it is monotone increasing towards \( U_i(z) \) as \( \lambda \to 0 \). This implies that the normalising constant

\[
Z_i^\lambda = \int_{z} \exp \left( -U_i^\lambda(z) - \frac{\| z - \theta \|^2}{2\rho^2} \right) \mathrm{d}z.
\]
is monotone decreasing towards $Z_i = \int x \exp \left(-U_i(x) - \frac{\|x - \theta^i\|^2}{2\rho^2} \right) \, dx$ as $\lambda \to 0$ (by the monotone convergence theorem). Therefore we have for any fixed $\Lambda > 0$, $0 < \lambda < \Lambda$, we have

$$\mu_i^\lambda(z) = \frac{\exp \left(-U_i^\lambda(z) - \frac{\|x - \theta^i\|^2}{2\rho^2} \right)}{Z_i^\lambda} \leq \frac{\exp \left(-U_i^\lambda(z) - \frac{\|x - \theta^i\|^2}{2\rho^2} \right)}{Z_i}.$$ 

This means that for $\lambda < \Lambda$, we have

$$\|z - \theta^i\|^2 \mu_i(z) - \mu_i^\lambda(z) \leq \|z - \theta\| \left( \mu_i(z) - \frac{\exp \left(-U_i^\lambda(z) - \frac{\|z - \theta^i\|^2}{2\rho^2} \right)}{Z_i} \right)$$

Using the strong-convexity of $-\log \mu_i$, it follows that it has a unique minimizer which we denote by $z_i^\star$. In particular, we have

$$\int_{x \in \mathbb{R}^d} \|z - \theta\| \mu_i(z) \, dz \leq \mu_i(z_i^\star) \int_{x \in \mathbb{R}^d} \|z - \theta\| \exp \left(-\frac{1}{2\rho^2}(z - z_i^\star)^2 \right) \, dz < \infty,$$

and with the same argument we can also show that

$$\int_{x \in \mathbb{R}^d} \|z - \theta\| \left( \mu_i(z) - \frac{\exp \left(-U_i^\lambda(z) - \frac{\|z - \theta^i\|^2}{2\rho^2} \right)}{Z_i} \right) < \infty,$$

hence using the pointwise convergence \(82\) it follows from the dominated convergence theorem and the bound that \(84\) that $W_1(\mu_i^\lambda, \mu_i) \to 0$ as $\lambda \to 0$. The same also holds for $W_1(\mu_i^\lambda, \mu_i')$, so we can conclude using \(83\) and the triangle inequality

$$W_1(\mu_i, \mu_i') \leq W_1(\mu_i^\lambda, \mu_i') + W_1(\mu_i^\lambda, \mu_i) + W_1(\mu_i', \mu_i').$$

Now we are ready to prove Theorems \[4\] and \[5\].

**Proof of Theorem \[4\]** We show this bound by a coupling argument. Let $\theta, \theta' \in \mathbb{R}^d$ be two parameter values, and $\mu_i$ and $\mu_i'$ denote the posterior distribution of $z_i$ given $\theta$ and $\theta'$, respectively. For every $i \in [b]$, we define couplings $(Z_i, Z_i')$ such that these couplings are independent for different indices $i$, $Z_i \sim \mu_i$, $Z_i' \sim \mu_i'$, and $\mathbb{E}(\|Z_i - Z_i'\|) = W_1(\mu_i, \mu_i')$. This is called the optimal coupling, whose existence is proven for example in Theorem 4.1 of \[52\].

Since conditionally on $z_{1:b}, \theta$ has a Gaussian distribution with mean $\sum_{i=1}^b \frac{x_i}{b}$ and covariance $\frac{1}{b} I_d$, it follows that the Wasserstein distance of $P_{SGS}(\theta, \cdot)$ and $P_{SGS}(\theta', \cdot)$ can be upper bounded using this coupling as

$$W_1 \left( P_{SGS}(\theta, \cdot), P_{SGS}(\theta', \cdot) \right) \leq \frac{\sum_{i=1}^b \mathbb{E}(\|Z_i - Z_i'\|)}{b} = \frac{\sum_{i=1}^b W_1(\mu_i, \mu_i')}{b} \leq \frac{\sum_{i=1}^b \frac{1}{1 + \rho^2 \rho_i}}{b},$$

where the last step follows by Lemma \[9\] The results now follows by the definition of the coarse Ricci curvature, and Corollary 21 of \[41\].

**Proof of Theorem \[5\]** We use a similar coupling argument as in the previous proof. Let $\theta, \theta' \in \mathbb{R}^d$ be two parameter values, and $\mu_i$ and $\mu_i'$ denote the posterior distribution of $z_i$ given $\theta$ and $\theta'$, respectively. For every $i \in [b]$, we define couplings $(Z_i, Z_i')$ such that these couplings are independent for different
indices $i, Z_i \sim \mu_i, Z_i' \sim \mu_i'$, and they are optimally coupled, i.e.

$$\mathbb{E}(|Z_i - Z_i'|) = W_1(\mu_i, \mu_i') \leq \frac{1}{1 + \rho^2 m_i}||\theta - \theta'||,$$

where the last inequality follows by Lemma 9. Since

$$\tilde{\pi}_p(\theta | Z_{B_j}) = \mathcal{N}(\theta; \tilde{\mu}_\theta(Z_{B_j}), \Sigma_{\theta}^{B_j}),$$

and we can set the subsampled set $j$ to be the same for both $\theta$ and $\theta'$ in the coupling (since the probability $p_j$ is independent of $\theta$), it follows that only the differences in the means matter for the Wasserstein distance, and hence

$$W_1(\pi_{\text{SSGS}}(\theta, \cdot | B_j), \pi_{\text{SSGS}}(\theta', \cdot | B_j)) \leq \|\tilde{\mu}_\theta(Z_{B_j}) - \tilde{\mu}_\theta(Z_{B_j}')\|$$

using the fact that $A_i = I_d$ by Assumption (A6)

$$= \left(\sum_{i \in B_j} w_{B_j}^i\right)^{-1} \sum_{i \in B_j} w_{B_j}^i (Z_i - Z_i')$$

using the fact that $\sum_{i \in B_j} w_{B_j}^i = b$ by Assumption (H1)

$$\leq \|\theta - \theta'|| \cdot \frac{1}{b} \sum_{i \in B_j} w_{B_j}^i \frac{\theta}{1 + \rho^2 m_i},$$

thus by taking average over all of the realizations $j \in [r]$ with probabilities $p_j$, we obtain that

$$W_1(\pi_{\text{SSGS}}(\theta, \cdot), \pi_{\text{SSGS}}(\theta', \cdot)) \leq \|\theta - \theta'|| \cdot \frac{1}{b} \sum_{j=1}^{r} \sum_{i=1}^{b} w_{B_j}^i p_j \frac{1}{1 + \rho^2 m_i}$$

using Assumption (A6)

$$= \|\theta - \theta'|| \cdot \frac{1}{b} \sum_{i=1}^{b} \frac{1}{1 + \rho^2 m_i}$$

thus the curvature lower bound follows. The existence and uniqueness of the stationary distribution $\pi_{\text{SSGS}}^\ast$ and the convergence bound (79) follows by Corollary 21 of [41].

\[\square\]

**D.2 Complexity bounds for implementing SGS by rejection sampling**

The following bound is a standard result in rejection sampling (see e.g. Section 2.3 of [46]).

**Lemma 10.** Suppose that $\mu(z) = \tilde{\mu}(z) / \tilde{Z}$ is the target distribution on $\mathbb{R}^d$, and $\nu(z)$ is the proposal distribution (both absolutely continuous w.r.t. the Lebesgue measure). Here $\tilde{\mu}(z)$ is the unnormalised target and $\tilde{Z}$ is the normalising constant (which is typically unknown). Suppose that the condition

$$\tilde{\mu}(z) \leq M \nu(z)$$

holds for some constant $M < \infty$ for every $z \in \mathbb{R}^d$. Under this assumption, if we take samples $Z_1, Z_2, \ldots$ from $\nu$ and accept $Z_i$ with probability $\frac{\tilde{\mu}(Z_i)}{M \nu(Z_i)}$, then the accepted samples will be distributed according to $\mu$. Moreover, the expected number of samples taken until the first acceptance is equal to $M / \tilde{Z}$.

The following lemma gives a complexity bound for rejection sampling for log-concave distributions. We assume that we have access to an approximation of the minimum of the strongly convex and smooth potential $U$, which will be denoted by $\tilde{U}$. The quality of this approximation is taken into account in the proposal distribution using the norm of $\nabla U(\tilde{z})$. 

27
Lemma 11 (An upper bound for rejection sampling for log-concave densities). Suppose that $\mu(z) \propto \exp(-U(z))$ is a distribution on $\mathbb{R}^d$ satisfying that $U$ is twice differentiable, and

$$ A I_d \preceq \nabla^2 U(z) \preceq B I_d $$

(86)

for some $0 < A \leq B$ (strongly convex and smooth). Let $z^*$ be the unique minimizer of $U$, $\tilde{z}$ another point (an approximation of $z^*$), and $\nu(z) = \mathcal{N}(z; \tilde{z}, \tilde{A}^{-1} I_d)$, where

$$ \tilde{A} = A + \frac{\|\nabla U(\tilde{z})\|^2}{2d} - \frac{\|\nabla^2 U(z)\|}{4d^2} + \frac{A\|\nabla U(\tilde{z})\|^2}{d}. $$

(87)

Suppose that we take samples $Z_1, Z_2, \ldots$ from $\nu$, and accept them with probability

$$ \mathbb{P}(Z_j \text{ is accepted}) = \exp \left( -\frac{\|\nabla U(z)^2\|}{2(A - A)} - \frac{U(z) - U(\tilde{z})}{2} + \frac{\tilde{A}\|z - \tilde{z}\|^2}{2} \right). $$

Then these accepted samples are distributed according to $\mu$. Moreover, the expected number of samples taken until one is accepted is less than or equal to $\left( B/\tilde{A} \right)^{d/2} \exp \left[ \frac{\|\nabla U(\tilde{z})\|^2}{2(A - A)} \left( \frac{1}{A - A} - \frac{1}{B} \right) \right]$.

**Proof.** The proposal distribution equals

$$ \nu(z) = \mathcal{N}(z; \tilde{z}, \tilde{A}^{-1} I_d) $$

$$ = \exp \left( -\frac{\tilde{A}\|z - \tilde{z}\|^2}{2} \right) \cdot \left( \frac{\tilde{A}}{2\pi} \right)^{d/2}. $$

We define the unnormalised version of $\mu$ as

$$ \tilde{\mu}(z) = \exp(-U(z) - U(\tilde{z})) \cdot \left( \frac{\tilde{A}}{2\pi} \right)^{d/2}. $$

Notice that

$$ U(z) - U(\tilde{z}) = \left\langle \int_{t=0}^{1} \nabla U(\tilde{z} + t(z - \tilde{z})) \, dt, z - \tilde{z} \right\rangle; $$

by the intermediate value theorem, there is some $\tilde{z}(t)$ such that

$$ = \left\langle \nabla U(\tilde{z}), z - \tilde{z} \right\rangle + \left( z - \tilde{z}, \left( \int_{t=0}^{1} t \nabla^2 U(z(t)) \, dt \right)^T (z - \tilde{z}) \right\rangle; $$

using the assumption (86) it follows

$$ \geq -\|\nabla U(\tilde{z})\|\|z - \tilde{z}\| + \frac{A}{2} \|z - \tilde{z}\|^2. $$

Based on this, it follows that

$$ \frac{\tilde{\mu}(z)}{\nu(z)} \leq \exp \left( -\frac{\|\nabla U(\tilde{z})\|\|z - \tilde{z}\| - \frac{A}{2} \|z - \tilde{z}\|^2}{2(A - A)} \right) \leq \exp \left( \frac{\|\nabla U(\tilde{z})\|^2}{2(A - A)} \right), $$

hence we have $\tilde{\mu}(z) \leq M \nu(z)$ for $M = \exp \left( \frac{\|\nabla U(\tilde{z})\|^2}{2(A - A)} \right)$.

For the normalising constant, we have

$$ \tilde{Z} = \int_{z \in \mathbb{R}^d} \tilde{\mu}(z) \, dz = \exp(U(\tilde{z}) - U(z^*)) \cdot \left( \frac{\tilde{A}}{2\pi} \right)^{d/2} \cdot \int_{z \in \mathbb{R}^d} \exp(-U(z) - U(z^*)) \, dz $$

28
where in the last step we have used the fact that for \( z \)

The parameter \( \tilde{\theta} \) that Assumptions (86)

Only the solution with the unique minimum is taken at a point where the derivative is zero. This point, denoted by \( \tilde{\theta} \), is easy to check that this is a strictly convex function of \( \tilde{\theta} \), hence by rearrangement

hence by rearrangement

\[
(\tilde{\theta} - A)^2 - (\|\nabla U(\tilde{\theta})\|^2/d)\tilde{\theta} = 0 \\
\tilde{\theta}^2 - (2A + \|\nabla U(\tilde{\theta})\|^2/d)\tilde{\theta} + A^2 = 0 \\
\tilde{\theta} = \frac{(2A + \|\nabla U(\tilde{\theta})\|^2/d) \pm \sqrt{(2A + \|\nabla U(\tilde{\theta})\|^2/d)^2 - 4A^2}}{2} \\
= A + \|\nabla U(\tilde{\theta})\|^2/(2d) \pm \sqrt{\|\nabla U(\tilde{\theta})\|^4/(4d^2) + 2A\|\nabla U(\tilde{\theta})\|^2/d}.
\]

Only the solution with the \(-\) sign falls in the interval \((0, A)\), hence it is the minimizer of \( M/\tilde{Z} \). \( \square \)

**Corollary 2** (Computational complexity of rejection sampling for sampling \( \theta \) given \( \theta \)). Suppose that Assumptions \( (A_2) \) and \( (A_3) \) (smoothness and convexity) hold, and that \( U_i \) is \( m_i \)-strongly convex
We stop once the condition \( \| \nabla \theta_i(z_i) \|^2 / 2 \rho^2 \leq \| A_i \theta_i - z_i \|^2 / 2 \rho^2 \),

which is satisfied, and set \( \tilde{z}_i = z_i \).

\( \text{Suppose that we take samples } Z_1, Z_2, \ldots \text{ from } V_i, \text{ and } \hat{z}_i(\theta) \text{ be another point (an approximation of } z_i(\theta)) \).

Let \( G \) be the approximate minimizer of \( V_i \) in each iteration, it follows that we need \( i = 1 + \frac{1}{2} \sqrt{1 / \rho^2 + m_i} \) iterations before stopping.

Then these accepted samples are distributed according to \( \pi_\rho(z_i) \). Moreover, the expected number of samples taken until one is accepted is equal to

\[ E_i := \left( \frac{1}{\rho^2 + M_i} \right)^{d_i/2} \cdot \exp \left[ \frac{1}{2} \left( \frac{1}{\rho^2 + m_i - A_i} - \frac{1}{1 / \rho^2 + M_i} \right) \right], \]

which is less than or equal to 2 if

\[ \rho^2 (2 d_i (M_i - m_i) - m_i) \leq 1 \quad \text{and} \quad \frac{1}{\rho^2} \frac{\sqrt{1 / \rho^2 + m_i}}{\sqrt{d_i}}. \]

\[ 2 \cdot \frac{1}{\rho^2} \frac{\sqrt{1 / \rho^2 + m_i}}{\sqrt{d_i}} \]

is satisfied, and set \( z_i \) to \( z_i \).

Since the condition number of the function \( V_i \) equals \( \kappa_i = \frac{1 + \rho^2 M_i}{\rho^2 + M_i} \), and the gradient descent decreases the norm of the gradient by a factor of \( 1 - 1 / \kappa_i \) in each iteration, it follows that we need at most

\[ \log \left( \frac{\| V_i(A, \theta) \|}{\log(1/(1 - 1 / \kappa_i))} \right) \]

iterations before stopping.

\[ \text{Proof of Corollary 3} \]

The fact that the accepted samples are distributed according to \( \pi_\rho(z_i) \) and the formula (90) about the expected number of samples until acceptance follows from Lemma 11.

Let \( G := \| V_i(\tilde{z}_i) \| \), then \( \hat{A}_i = 1 / \rho^2 + m_i + G^2 / (2 d_i) - \sqrt{G^4 / (4 d_i^2) + G^2 (1 / \rho^2 + m_i) / d_i} \), and we have

\[ \log(E_i) = \frac{d_i}{2} \log \left( \frac{1 / \rho^2 + M_i}{(1 / \rho^2 + m_i + G^2 / (2 d_i) - \sqrt{G^4 / (4 d_i^2) + G^2 (1 / \rho^2 + m_i) / d_i})} \right) \]

(92)

\[ + \frac{G^2}{2} \left( \frac{1}{\sqrt{G^4 / (4 d_i^2) + G^2 (1 / \rho^2 + m_i) / d_i} - G^2 / (2 d_i) - \frac{1}{1 / \rho^2 + M_i}} \right). \]

(93)
For the first part, notice that
\[
\log \left( \frac{1/\rho^2 + M_i}{1/\rho^2 + m_i + G^2/(2d_i)} - \sqrt{G^4/(4d_i^2) + G^2 (1/\rho^2 + m_i) / d_i} \right)
\]
\[
= \log \left( \frac{1/\rho^2 + M_i}{1/\rho^2 + m_i} \right) + \log \left( \frac{1/\rho^2 + m_i}{1/\rho^2 + m_i + G^2/(2d_i)} - \sqrt{G^4/(4d_i^2) + G^2 (1/\rho^2 + m_i) / d_i} \right)
\]
\[
= \log \left( 1 + \frac{\rho^2(M_i - m_i)}{1 + \rho^2 m_i} \right) + \log \left( \frac{1}{1 + c - \sqrt{c^2 + 2c}} \right),
\]
where \( c = G^2/(2d_i) \). Now using the fact that \( \log(1+x) \leq x \) for \( x > 0 \), and that \( \log \left( \frac{1}{1+c-\sqrt{c^2+2c}} \right) \leq \sqrt{2c} \) for \( c \geq 0 \), it follows that we have
\[
d_i \log \left( \frac{1/\rho^2 + M_i}{1/\rho^2 + m_i + G^2/(2d_i)} - \sqrt{G^4/(4d_i^2) + G^2 (1/\rho^2 + m_i) / d_i} \right)
\]
\[
\leq \frac{d_i}{2} \left( \frac{\rho^2(M_i - m_i)}{1 + \rho^2 m_i} + \frac{G}{\sqrt{d_i(1/\rho^2 + m_i)}} \right).
\]
For the second part \([93]\),
\[
\frac{G^2}{2} \left( \frac{1}{\sqrt{G^4/(4d_i^2) + G^2 (1/\rho^2 + m_i) / d_i}} - \frac{1}{1/\rho^2 + M_i} \right)
\]
\[
= \frac{d_i}{2} \left( \frac{1}{1 + 4 (1/\rho^2 + m_i) d_i/G^2 - 1} - \frac{G^2}{2} \cdot \frac{1}{1/\rho^2 + M_i} \right)
\]
using the fact that \( \frac{1}{\sqrt{1+x}} \leq \frac{2}{\sqrt{x}} \) for \( x \geq 2 \), for \( G \leq \sqrt{2d_i(1/\rho^2 + m_i)} \), we have
\[
\leq G \cdot \frac{\sqrt{d_i}}{\sqrt{1/\rho^2 + m_i}} = \frac{G^2}{2} \cdot \frac{1}{1/\rho^2 + M_i}.
\]
Hence by combining these terms, we obtain that for \( G \leq \sqrt{2d_i(1/\rho^2 + m_i)} \),
\[
\log(E_i) \leq \frac{d_i}{2} \frac{\rho^2(M_i - m_i)}{1 + \rho^2 m_i} + G \cdot \frac{3}{2} \cdot \frac{\sqrt{d_i}}{\sqrt{(1/\rho^2 + m_i)}} - \frac{G^2}{2} \cdot \frac{1}{1/\rho^2 + M_i}, \quad (94)
\]
Under the first part of assumption \([91]\), \( \rho^2(2d_i(M_i - m_i) - m_i) \leq 1 \), one can check that \( \frac{d_i}{2} \frac{\rho^2(M_i - m_i)}{1 + \rho^2 m_i} \leq \frac{1}{4} \). Using the second part of \([91]\), \( G \leq \frac{3}{2} \cdot \frac{\sqrt{1/\rho^2 + m_i}}{\sqrt{d_i}} \), it follows that \( G \cdot \frac{3}{2} \cdot \frac{\sqrt{d_i}}{\sqrt{(1/\rho^2 + m_i)}} - \frac{G^2}{2} \cdot \frac{1}{1/\rho^2 + M_i} \leq \log(2) - \frac{1}{4} \), so \( \log(E_i) \leq \log(2) \) and our claim holds. \( \square \)

**D.3 Computational complexity bounds in Wasserstein distance**

By combining our convergence and approximation bounds for the single splitting case, we obtain the following complexity result.

**Proposition 2** (Computational complexity bound in Wasserstein distance). Suppose that \( b = 1 \) (single splitting case), and that Assumptions \((A_2), (A_3) \) and \((A_4)\) (strong convexity, smoothness and \( A_1 = I_d \)) hold. Let \( \theta^* \) be the unique minimizer of \( U_1(\theta) \). Suppose that \( \epsilon \leq 1 \). Then with the choice
\[
\rho^2 = \max \left( \frac{\epsilon^2}{4m_1}, \frac{\epsilon}{\sqrt{m_1 M_1}} \right), \quad (95)
\]
and number of iterations
\[ t(\epsilon) \geq \frac{\log \left( \frac{3}{2} \right)}{\log \left( 1 + \max \left( \frac{\epsilon^2}{\epsilon^2}, \epsilon \sqrt{\frac{m_1}{M}} \right) \right)}, \tag{96} \]
we have
\[ W_1(P_t^{(\epsilon)}(\theta^*, \cdot), \pi) \leq \frac{\epsilon}{\sqrt{m_1}} \sqrt{d}, \]
i.e. the SGS with parameter \( \rho \) defined as \( (95) \) initialised at \( \theta^* \) will be less than \( \frac{\epsilon}{\sqrt{m_1}} \) Wasserstein distance from the target \( \pi \) after \( t(\epsilon) \) steps.

Remark 3. On Table 2 we compare our bounds for SGS with single splitting implemented based on rejection sampling (see Corollary 2) with the existing Wasserstein convergence rates in the literature. We are competitive in the \( 0 < \epsilon < \frac{1}{d\sqrt{\kappa}} \) precision range (where rejection sampling can be efficiently implemented).

Proof of Proposition 2. From Theorem 3 it follows that if \( \rho \) is chosen as in \( (95) \), then
\[ W_1(\pi_{\rho}, \pi) \leq \frac{\epsilon}{2} \cdot \frac{\sqrt{d}}{\sqrt{m_1}}. \tag{97} \]
From Proposition 1 part (ii) in [23] it follows that for the initial distribution \( \delta_{\theta^*} \) (Dirac-\( \delta \) centered at \( \theta^* \)), we have
\[ W_1(\delta_{\theta^*}, \pi) \leq W_2(\delta_{\theta^*}, \pi) \leq \frac{\sqrt{d}}{\sqrt{m_1}}, \]
and hence by combining this with \( (97) \) using the triangle inequality and the assumption \( \epsilon \leq 1 \), it follows that
\[ W_1(\delta_{\theta^*}, \pi_{\rho}) \leq \frac{3}{2} \cdot \frac{\sqrt{d}}{\sqrt{m_1}}. \]
Now from Theorem 4 it follows that the coarse Ricci curvature of SGS is lower bounded by
\[ K_{SGS} := \frac{\rho^2 m_1}{1 + \rho^2 m_1}, \]
and therefore by Corollary 21 of [41], we have
\[ W_1(P_t^{(\epsilon)}(\theta^*, \cdot), \pi_{\rho}) \leq W_1(\delta_{\theta^*}, \pi_{\rho}) \cdot (1 - K_{SGS})^{t(\epsilon)} \leq \frac{\epsilon}{2} \cdot \frac{\sqrt{d}}{\sqrt{m_1}}. \]
The claim of the theorem now follows by the triangle inequality.

D.4 Computational complexity bounds in total variation distance

Our next result shows some convergence results in total variation distance.

**Proposition 3** (Computational complexity bound in TV distance). Suppose that Assumptions \( (A_2), (A_3), (A_4), (A_5) \) and \( (A_6) \) hold. Let \( \theta^* \) be the unique minimizer of \( U(\theta) = \sum_{i=1}^b U_i(\theta) \), and let \( \nu(\theta) := \mathcal{N}(\theta; \theta^*, \left( \sum_{i=1}^b M_i \right)^{-1} I_d) \) be the initial distribution. Then for any \( 0 < \epsilon \leq 1 \), with the choice
\[ \rho^2 \leq \frac{d \sum_{i=1}^b M_i \left( \sqrt{1 + 2\epsilon \left( \frac{d \sum_{i=1}^b M_i}{\sum_{i=1}^b M_i} - 1 \right)} \right)^2}{\left( \frac{4 \sum_{i=1}^b M_i M_i}{\sum_{i=1}^b M_i M_i} \right)^2}, \tag{98} \]
and number of iterations
\[ t(\epsilon) \geq \frac{\log \left( \frac{3}{2} \right) + C/2}{(\rho^2/b) \sum_{i=1}^b \frac{m_i}{1 + m_i \rho^2}}, \tag{99} \]
Table 5: Comparison of convergence rates in TV distance with the literature, starting from a Gaussian distribution centered at the minimizer $\theta^*$ of the $m$-strongly convex and $M$-smooth potential $U(\theta)$, with condition number $\kappa = \frac{M}{m}$. SGS with single splitting is implemented based on rejection sampling.

$O^*(\cdot)$ denotes $O(\cdot)$ up to polylogarithmic factors, $\nu_{m}(\theta) = N(\theta; \theta^*, \frac{I}{m})$ and $\nu_{M}(\theta) = N(\theta; \theta^*, \frac{I}{M})$.

| Ref. | Method                        | Condition on $\epsilon$ | Grad/func evals for TV err. $\epsilon$ |
|------|-------------------------------|--------------------------|----------------------------------------|
| [22] | ULA started from $\theta^*$   | $0 \leq \epsilon \leq 1$ | $O^*(\kappa^2d^2/\epsilon^2)$          |
| [14], [21] | ULA started from $\nu_{m}$ | $0 < \epsilon \leq 1$    | $O^*(\kappa^2d^2/\epsilon^2)$          |
| [25] | MALA started from $\nu_{M}$  | $0 < \epsilon \leq 1$    | $O(\kappa^2d^2\log^{1.5}\left(\frac{\kappa}{\epsilon^{1/d}}\right))$ |
| this paper | SGS started from $\nu_{M}$ | $0 < \epsilon \leq 1$    | $O^*(\kappa^2d^2/\epsilon)$          |

we have

$$\|\nu_{\text{SGS}}^{(\epsilon)} - \pi\|_{\text{TV}} \leq \epsilon.$$  

Here

$$C = \sum_{i=1}^{b} M_i \|\theta^* - \theta_i^*\|^2 + \frac{bd}{8} + \frac{d}{2} \log\left(\frac{2 \sum_{i=1}^{b} M_i}{\sum_{i=1}^{b} m_i}\right),$$

with $\theta_i^*$ denoting the unique minimizer of $U_i$.

This means that starting from $\nu$, after $t(\epsilon)$ SGS steps, we are at most $\epsilon$ TV distance from $\pi$.

**Remark 4.** In the single splitting case ($b = 1$), we have $M_1 \leq dM_1$ (based on the argument using integration by parts at the end of Theorem 3). Since $\sqrt{1 + x} - 1 \geq \frac{x}{2}$ for $0 \leq x \leq 2$, we can show that (101) is satisfied if $\rho^2 = \frac{\epsilon}{8dM_1}$, and hence

$$t(\epsilon) \geq \frac{1}{\epsilon} \log \left(\frac{2}{\rho^2}\right) d + \frac{d}{2} \log\left(2 + \frac{5}{\rho^2}\right) + 5 \log\left(\frac{M_1}{m_1}\right) \cdot \frac{M_1}{m_1}.$$  

SGS steps starting from $\nu$ suffice for a sample with at most $\epsilon$ TV distance from $\pi$. Table 5 compares this method implemented using rejection sampling (see Corollary 2) with existing results in the literature. As we can see, compared to ULA we have a better dependence on $\epsilon$ and $\kappa$, but worse dependence on $d$ (hence for sufficiently small $\epsilon$, we get better rates). However, MALA seem to have better convergence rates in total variation distance in general, except in badly conditioned situations, where the rates for SGS can be better.

The next two lemmas will be used for obtaining our total variation distance convergence rates.

**Lemma 12.** Suppose that $U : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and $M$-gradient-Lipschitz. Then for every $x \in \mathbb{R}^d$, we have

$$\|\nabla U(x)\|^2 \leq 2M(U(x) - \inf_{\mathbb{R}^d} U(x)).$$

**Proof.** Let $x' = x - \nabla U(x)/M$, then we have

$$U(x) - U(x') = \int_{t=0}^{1} \langle \nabla U(x + t(x' - x)), x - x' \rangle \, dt$$

$$= \langle \nabla U(x), x - x' \rangle + \int_{t=0}^{1} \langle \nabla U(x + t(x' - x)) - \nabla U(x), x - x' \rangle \, dt$$

using the $M$-gradient Lipschitz property

$$\geq \langle \nabla U(x), x - x' \rangle + \int_{t=0}^{1} Mt\|x - x'\|^2 \, dt.$$
With an analogous argument using (103), we obtain that

\[
\frac{\|\nabla U(x)\|^2}{2M},
\]

hence the result. \(\square\)

**Lemma 13.** Suppose that Assumptions (A2), (A5) and (A6) hold. Let \(\theta^*\) be the minimizer of \(U(\theta) = \sum_{i=1}^b U_i(\theta), A^*_i = \text{the minimizer of } U_i(\theta), \text{and } \nu(\theta) = N(\theta; \theta^*, (\sum_{i=1}^b M_i)^{-1} I_d).\) Suppose that \(\rho^2 \leq \frac{1}{\max_i \{\rho^2; M_i\}}\). Then \(\frac{\nu(\theta)}{\pi(\theta)} \leq C_\rho\) for every \(\theta \in \mathbb{R}^d\), where

\[
C_\rho := \exp \left( \sum_{i=1}^b \frac{M_i \|\theta^* - \theta^*_i\|^2}{2} \right) \cdot \prod_{i=1}^b \left(1 + \rho^2 M_i\right)^{d/2} \cdot \left(\frac{2 \sum_{i=1}^b M_i}{\sum_{i=1}^b m_i}\right)^{d/2}. \tag{101}
\]

**Proof.** Let

\[
U_i^\rho(\theta) := -\log \int_{z_i \in \mathbb{R}^d} \exp \left(-U_i(z_i) - \frac{\|z_i - \theta\|^2}{2\rho^2}\right) \cdot \frac{dz_i}{(2\pi \rho^2)^d/2}.
\]

Then using the conditional independence of \(z_i\) given \(\theta\), we can see that the unnormalised negative log-likelihood of \(\pi_\rho(\theta)\) can be written as

\[
U_\rho(\theta) := \sum_{i=1}^b U_i^\rho(\theta).
\]

With this, we are able to write

\[
\pi_\rho(\theta) = \frac{\exp(-U_\rho(\theta))}{Z_\rho},
\]

for a normalising constant \(Z_\rho\). Using the assumptions (A2) and (A5) (strong convexity and smoothness), we have

\[
U_i(z_i) \geq U_i(\theta) + \langle \nabla U_i(\theta), z_i - \theta \rangle + \frac{m_i}{2} \|z_i - \theta\|^2 \tag{102}
\]

\[
U_i(z_i) \leq U_i(\theta) + \langle \nabla U_i(\theta), z_i - \theta \rangle + \frac{M_i}{2} \|z_i - \theta\|^2. \tag{103}
\]

Using (102), we have

\[
\exp(-U_i^\rho(\theta)) = \int_{z_i \in \mathbb{R}^d} \exp \left(-U_i(z_i) - \frac{\|z_i - \theta\|^2}{2\rho^2}\right) \cdot \frac{dz_i}{(2\pi \rho^2)^d/2}
\]

\[
\leq \exp(-U_i(\theta)) \int_{z_i \in \mathbb{R}^d} \exp \left(-\langle \nabla U_i(\theta), z_i - \theta \rangle - \left(\frac{1 + m_i \rho^2}{2\rho^2}\right) \|z_i - \theta\|^2\right) \cdot \frac{dz_i}{(2\pi \rho^2)^d/2}
\]

\[
= \exp(-U_i(\theta)) \int_{z_i \in \mathbb{R}^d} \exp \left(-\langle \nabla U_i(\theta), z_i \rangle - \left(\frac{1 + m_i \rho^2}{2\rho^2}\right) \|z_i\|^2\right) \cdot \frac{dz_i}{(2\pi \rho^2)^d/2}
\]

\[
= \frac{\exp(-U_i(\theta))}{(2\pi \rho^2)^d/2} \int_{z_i \in \mathbb{R}^d} \exp \left(-\left(\frac{1 + m_i \rho^2}{2\rho^2}\right) \|z_i\|^2 + \frac{\rho^2 \|\nabla U_i(\theta)\|^2}{2(1 + \rho^2 m_i)}\right) dz_i
\]

\[
\leq \exp(-U_i(\theta)) \cdot \frac{\rho^2 \|\nabla U_i(\theta)\|^2}{2(1 + \rho^2 m_i)^{d/2}} \tag{104}
\]

With an analogous argument using (103), we obtain that

\[
\exp(-U_i^\rho(\theta)) \geq \frac{\exp(-U_i(\theta))}{(1 + \rho^2 M_i)^{d/2}} \cdot \exp \left(\frac{\rho^2 \|\nabla U_i(\theta)\|^2}{2(1 + \rho^2 M_i)}\right). \tag{105}
\]
Let $U(\theta) = \sum_{i=1}^{b} U_i(\theta)$, then from (105), it follows that

$$\pi_\rho(\theta) = \frac{\exp(-U^\rho(\theta))}{Z_\rho} \geq \frac{\exp(-U(\theta))}{Z_\rho} \cdot \frac{1}{\prod_{i=1}^{b} (1 + \rho^2 M_i)^{d/2}} \cdot \frac{\exp\left(-U(\theta^*) - \sum_{i=1}^{b} M_i \|\theta - \theta^*_i\|^2\right)}{\prod_{i=1}^{b} (1 + \rho^2 M_i)^{d/2}}. \tag{106}$$

Using (104), we can upper bound $Z_\rho$ as

$$Z_\rho = \int_{\theta \in \mathbb{R}^d} \exp(-U^\rho(\theta)) d\theta \leq \int_{\theta \in \mathbb{R}^d} \exp\left(-U(\theta)\right) \exp\left(\frac{\rho^2 \sum_{i=1}^{b} \|\nabla U_i(\theta)\|^2}{2}\right) d\theta.$$

Now using Lemma 12, we have $\|\nabla U_i(\theta)\|^2 \leq 2M_i(U_i(\theta) - U_i(\theta^*_i))$, and hence

$$-U(\theta) + \rho^2 \sum_{i=1}^{b} \|\nabla U_i(\theta)\|^2 \leq -U(\theta^*) - \frac{1}{2} \sum_{i=1}^{b} (U_i(\theta) - U_i(\theta^*_i)) + \sum_{i=1}^{b} (U_i(\theta^*) - U_i(\theta^*_i))$$

$$= -U(\theta^*) - \frac{1}{2} \sum_{i=1}^{b} (U_i(\theta) - U_i(\theta^*_i)) + \sum_{i=1}^{b} (U_i(\theta^*) - U_i(\theta^*_i))$$

$$\leq -U(\theta^*) - \frac{1}{4} \sum_{i=1}^{b} M_i \|\theta - \theta^*_i\|^2 - \frac{1}{4} \sum_{i=1}^{b} m_i \|\theta - \theta^*_i\|^2.$$

Now by convexity of the $\|\cdot\|^2$, we have

$$\sum_{i=1}^{b} m_i \|\theta - \theta^*_i\|^2 = \left(\sum_{i=1}^{b} m_i\right) \sum_{i=1}^{b} \frac{m_i}{\sum_{i=1}^{b} m_i} \|\theta - \theta^*_i\|^2 \geq \left(\sum_{i=1}^{b} m_i\right) \left\|\theta - \sum_{i=1}^{b} \frac{m_i \theta^*_i}{\sum_{i=1}^{b} m_i}\right\|^2,$$

so we obtain that

$$Z_\rho \leq \int_{\theta \in \mathbb{R}^d} \exp(-U(\theta)) \exp\left(\frac{\rho^2 \sum_{i=1}^{b} \|\nabla U_i(\theta)\|^2}{2}\right) d\theta \leq \exp\left(-U(\theta^*) + \sum_{i=1}^{b} M_i \frac{\|\theta^* - \theta^*_i\|^2}{2}\right).$$
The following proposition is well known in the MCMC literature (but we have only found a proof for
Proposition 4 (Total variation convergence bound for reversible chains with positive spectral gap).
Suppose that a $P(z,\cdot)$ is a reversible Markov kernel on a Polish state space $\Omega$ with absolute spectral
gap $\gamma^* > 0$, and unique stationary distribution $\pi$. Then for any initial distribution $\nu$ that is absolutely
continuous with respect to $\pi$, and any number of steps $t \in \mathbb{Z}_+$, we have
\[
\|\nu P^t - \pi\|_{TV} \leq \frac{1}{2} \left( \mathbb{E}_\pi \left( \frac{d\nu}{d\pi} \right)^2 - 1 \right)^{1/2} \cdot (1 - \gamma^*)^t.
\]
Our proof is based on the following lemma.

Lemma 14. Suppose that $Q(x,dy)$ is a reversible Markov kernel on a Polish state space $\Omega$ with
stationary distribution $\pi$. Then for any distribution $\nu$ that is absolutely continuous with respect to $\pi,$
$\nu Q$ is also absolutely continuous with respect to $\pi,$ and for $\pi$-almost every $x \in \Omega,$ we have
\[
\frac{d(\nu Q)}{d\pi}(x) = \left( Q \left( \frac{d\nu}{d\pi} \right) \right)(x).
\]

Proof. The claim of the lemma is equivalent to showing that for every bounded measurable function
$f : \Omega \to \mathbb{R},$ we have
\[
\int_{x \in \Omega} \frac{d(\nu Q)}{d\pi}(x)f(x)\pi(dx) = \int_{x \in \Omega} \left( Q \left( \frac{d\nu}{d\pi} \right) \right)(x)f(x)\pi(dx).
\]
Since if we add a constant to $f,$ both sides increase by this constant, we can assume without loss of
generality that $f$ is non-negative. Under this assumption, we have
\[
\int_{x \in \Omega} \frac{d(\nu Q)}{d\pi}(x)f(x)\pi(dx) = \int_{x \in \Omega} f(x)(\nu Q)(dx)
\]
Now we are ready to prove our convergence bound in total variation distance.

Using Lemma 14 with $\Pi$ the linear operator

Proof of Proposition 4. We define the Hilbert space $L^2(\pi)$ as measurable functions $f$ on $\Omega$ satisfying that $\mathbb{E}_\pi(f^2) < \infty$, endowed with the scalar product $\langle f, g \rangle_{\pi} = \int_{\pi} f(z)g(z)\pi(\text{d}z)$. Let us define the linear operator $\Pi(f)(z) := \mathbb{E}_\pi(f)$ for any $f \in L^2(\pi), z \in \Omega$.

Using Lemma 14 with $Q = P^t$, it follows that

$$\|\nu P^t - \pi\|_{\text{TV}} = \frac{1}{2} \int_{\pi} \left| \frac{\text{d}\nu P^t}{\text{d}\pi}(x) - 1 \right| \pi(\text{d}x)$$

using the Cauchy-Schwarz inequality and the fact that $\int_{\pi} \pi(\text{d}x) = 1$, we have

$$\leq \frac{1}{2} \sqrt{\int_{\pi} \left( \frac{\text{d}\nu P^t}{\text{d}\pi}(x) - 1 \right)^2 \pi(\text{d}x).}$$

Using Lemma 14 again, the integral inside the square root can be further bounded as

$$= \int_{\pi} \left( (P^t - \Pi) \left( \frac{\text{d}\nu}{\text{d}\pi} \right) (x) \right)^2 \pi(\text{d}x) = \int_{\pi} \left( (P - \Pi)^t \left( \frac{\text{d}\nu}{\text{d}\pi} \right) (x) \right)^2 \pi(\text{d}x)$$

$$= \int_{\pi} \left( \min_{x,y} \|P - \Pi\|_{\pi}^2 \left\| \frac{\text{d}\nu}{\text{d}\pi} - 1 \right\|_{\pi}^2 = (1 - \gamma^*)^{2t} \left\| \frac{\text{d}\nu}{\text{d}\pi} - 1 \right\|_{\pi}^2$$

and the claim of the proposition follows by substituting this into (108). 

Now we are ready to prove our convergence bound in total variation distance.
Proof of Proposition 3. From (58), we have
\[ \| \pi_\rho - \pi \|_{TV} \leq \rho^2 \sum_{i=1}^{b} d_i M_i + \rho^4 \sum_{i=1}^{b} M_i M_i. \]

Then, a sufficient condition to satisfy \( \| \pi_\rho - \pi \|_{TV} \leq \epsilon / 2 \) is to have
\[
\rho^2 \sum_{i=1}^{b} d_i M_i + \rho^4 \sum_{i=1}^{b} M_i M_i \leq \frac{\epsilon}{2}
\]
\[
\rho^4 \sum_{i=1}^{b} M_i M_i + \rho^2 \sum_{i=1}^{b} d_i M_i - \frac{\epsilon}{2} \leq 0
\]
\[
R^2 \sum_{i=1}^{b} M_i M_i + R \sum_{i=1}^{b} d_i M_i - \frac{\epsilon}{2} \leq 0, \text{ with } R = \rho^2.
\]

This inequality is satisfied under the condition (98).

From Corollary 1, we know that the absolute spectral gap of SGS satisfies that \( \gamma^* \geq K_{\text{SGS}} \) (defined in (73)), and Proposition 4 implies that
\[
\| \nu P_{\text{SGS}}^{t(\epsilon)} \pi_\rho \nu \|_{TV} \leq \sqrt{\mathbb{E}_\nu \left[ \left( \frac{d \nu}{d \pi_\rho} \right)^2 \right] - 1} \cdot (1 - \gamma^* t(\epsilon))
\]
\[
\leq \sqrt{\mathbb{E}_\nu \left( \frac{d \nu}{d \pi_\rho} \right) \cdot (1 - K_{\text{SGS}} t(\epsilon))}
\]
\[
\leq \sqrt{C_{\rho}(1 - K_{\text{SGS}})^t},
\]

where in the last step we have used Lemma 13 \( (C_{\rho} \text{ is defined as in (101))}. \) Using the fact that \( \sqrt{1 + x} \leq 1 + x/2 \) for \( x \geq 0 \), it is easy to check that the condition \( \rho^2 \leq \frac{1}{4 \max_{1 \leq i \leq b} M_i} \) of Lemma 13 is satisfied under the condition (98). By some algebra, using the definition of \( t(\epsilon) \), and the fact that \( \frac{1}{\mathbb{E}_\nu(t(1-x))} \leq \frac{1}{2} \) for \( 0 < x < 1 \), the above bound implies that
\[
\| \nu P_{\text{SGS}}^{t(\epsilon)} \pi_\rho \|_{TV} \leq \frac{\epsilon}{2},
\]

and hence the claim of the Proposition follows by the triangle inequality.

\[ \square \]

D.5 Additional details for the toy Gaussian example

This section gives additional details concerning the results depicted on Figure 2. For each splitting strategy introduced in Section 3.2, we give explicit formulas for the bounds on both TV and 1-Wasserstein distances.

D.5.1 Splitting strategy 1

Starting from an initial value \( \theta_0 \sim \nu \), we now show the explicit form of the Markov transition kernel \( \nu P_{\text{SGS}}^t \) after \( t \) iterations. To this purpose, we take advantage that the \( \theta \)-chain corresponds in this case to an auto-regressive process of order 1. Indeed, the conditional distributions of \( \theta \) and \( z_{1:b} \) writing
\[ \pi_\rho(z_i|\theta) = \mathcal{N}(z_i; \mu \rho^2 + \theta \sigma^2 / \sigma^2 + \rho^2, \rho^2 \sigma^2 / \sigma^2 + \rho^2 + \sigma^2), \forall i \in [b] \]
\[ \pi_\rho(\theta|z_{1:b}) = \mathcal{N}(\theta; \bar{z}, \rho^2 / b), \text{ where } \bar{z} := \frac{1}{b} \sum_{i=1}^{b} z_i, \]
we have
\[ P_{SGS} \coloneqq \Pr \left( \theta^t \mid \theta^{t-1} \right) = \mathcal{N} \left( \theta^t ; \frac{\sigma^2}{\sigma^2 + \rho^2} \theta^{t-1} + \frac{\rho^2}{\sigma^2 + \rho^2} \mu, \frac{2\rho^2\sigma^2 + \rho^4}{b(\rho^2 + \sigma^2)} \right). \]

By a straightforward induction, it follows that the Markov transition kernel \( \nu P^t \) after \( t \) iterations and with initial distribution \( \nu \) has the form
\[ \nu P^t_{SGS} \coloneqq \Pr \left( \theta^t \mid \theta^0 \sim \nu \right) = \mathcal{N} \left( \theta^t ; \left( \frac{\sigma^2}{\sigma^2 + \rho^2} \right)^t \theta^0 + \frac{\rho^2\mu}{\sigma^2 + \rho^2} \sum_{i=0}^{t-1} \left( \frac{\sigma^2}{\sigma^2 + \rho^2} \right)^i \frac{2\rho^2\sigma^2 + \rho^4}{b(\rho^2 + \sigma^2)} \sum_{i=0}^{t-1} \left( \frac{\sigma^2}{\sigma^2 + \rho^2} \right)^i \right). \]

Straightforward calculus lead to the following closed-form expressions for the quantities appearing in the bounds of Section 4:
\[ K_{SGS} = \frac{\rho^2}{\rho^2 + \sigma^2} \]
\[ \log(C_{\rho}) = \frac{b}{2} \log \left( 1 + \frac{\rho^2}{\sigma^2} \right) + \frac{1}{2} \log \left( \frac{\sigma^2}{\sigma^2 - 2\rho^2} \right), \text{ for any } \rho < \sigma / \sqrt{2} \]
\[ M_i = m_i = \sigma^2, \forall i \in [b] \]
\[ \theta^* = \theta^*_i = \mu, \forall i \in [b]. \]

D.5.2 Splitting strategy 2

Similar calculus as in the above section can be undertaken by simply replacing \( \rho^2 \) by \( \rho^2 b \).

E Additional details on the logistic regression experiment in Section 5.1

E.1 Problem statement

We consider the observation of binary responses \( y \in \mathbb{R}^n \) assumed to stand for conditionally independent Bernoulli random variables with probability of success \( h (x_i^T \theta) \). The function \( h \) is the logistic link defined for all \( t \in \mathbb{R} \) as \( h(t) = \exp(t) / (1 + \exp(t)) \), \( x_i \) stands for the covariates associated to the \( i \)-th observation \( y_i \) and \( \theta \in \mathbb{R}^d \) represents the vector of regression coefficients to infer. We adopt a classical zero-mean Gaussian prior with precision \( 2\pi I_d \) on \( \theta \) leading to the posterior with density
\[ \pi(\theta \mid y) \propto \exp \left( -\tau \| \theta \|^2 - \sum_{i=1}^n U_i (x_i^T \theta) \right), \quad (110) \]
where \( U_i (u) = -y_i \log (h(u)) - (1 - y_i) \log (1 - h(u)) \).

Adopting a full splitting strategy \( (b = n + 1) \) leads to the joint probability density function
\[ \pi_\rho(\theta, z_{1:n} \mid y) \propto \exp \left( -\tau \| z_b \|^2 - \frac{1}{2\rho^2} \| \theta - z_b \|^2 - \frac{1}{2\rho^2} \| X \theta - z_{1:n} \|^2 - \sum_{i=1}^n U_i (z_i) \right), \quad (111) \]
where \( \rho > 0 \) and \( X \) is the matrix made of all covariates vectors.

E.2 Implementation details

This section details how the sampling from each conditional distribution has been conducted in our experiments.
E.2.1 Sampling the auxiliary vector $z_{1:n}$

The conditional distribution associated to the auxiliary variable $z_{i,i\in[n]}$ writes

$$
\pi_\rho(z_i | y_i, \theta) \propto \exp\left(-\frac{1}{2\rho^2} \left(\mathbf{x}_i^T \theta - z_i\right)^2 - U_i(z_i)\right).
$$

(112)

Since this distribution is univariate and log-concave, one can sample from it exactly by using adaptive rejection sampling (ARS) [26]. In order to have a sampling step more amenable to parallelization, we used instead the unadjusted Langevin algorithm (ULA) to sample from (112). In this case, the ULA writes

$$
z_i^{[t+1]} = z_i^{[t]} - \delta \nabla U_i(z_i^{[t]}) - \frac{\delta}{\rho^2} \left(z_i^{[t]} - \mathbf{x}_i^T \theta\right) + \sqrt{2\delta} \epsilon^{[t+1]}, \quad \epsilon^{[t+1]} \sim \mathcal{N}(0,1).
$$

(113)

The step size $\delta$ has been set to $L_1^{-1}$ where $L_1$ stands for the Lipschitz constant of $-\nabla \log \pi_\rho(z_i | y_i, \theta)$, that is $L_1 = 1/\rho^2 + 1/4$.

E.2.2 Sampling the auxiliary vector $z_b$

The conditional distribution associated to the auxiliary variable $z_b$ writes

$$
\pi_\rho(z_b | \theta) \propto \exp\left(-\tau \|z_b\|^2 - \frac{1}{2\rho^2} \|\theta - z_b\|^2\right).
$$

(114)

This distribution is a Gaussian distribution $\mathcal{N}(\mu_{z_b}, \Sigma_{z_b})$ where

$$
\Sigma_{z_b} = \frac{\rho^2}{2\tau \rho^2 + 1} \mathbf{I}_d
$$

(115)

$$
\mu_{z_b} = \frac{\theta}{2\tau \rho^2 + 1}.
$$

(116)

E.2.3 Sampling the parameter of interest

The conditional distribution associated to $\theta$ writes

$$
\pi_\rho(\theta | z_{1:n}) \propto \exp\left(-\frac{1}{2\rho^2} \|\theta - z_b\|^2 - \frac{1}{2\rho^2} \|\mathbf{X} \theta - z_{1:n}\|^2\right).
$$

(117)

This distribution is a Gaussian distribution $\mathcal{N}(\mu_{\theta}, \Sigma_{\theta})$ where

$$
\Sigma_{\theta} = \rho^2 \left(\mathbf{X}^T \mathbf{X} + \mathbf{I}_d\right)^{-1}
$$

(118)

$$
\mu_{\theta} = \left(\mathbf{X}^T \mathbf{X} + \mathbf{I}_d\right)^{-1} \left(z_b + \mathbf{X}^T z_{1:n}\right).
$$

(119)

Note that the matrix $\left(\mathbf{X}^T \mathbf{X} + \mathbf{I}_d\right)^{-1}$ can be pre-computed. In addition, it is worth pointing out that sampling from this Gaussian distribution can be achieved by only drawing univariate normal samples thanks to the exact perturbation-optimization (E-PO) algorithm of [42].

E.3 Comparison between ULA and rejection sampling

In this section, we compare the approximate sampling from the conditionals $\pi_\rho(z_i | \theta)$ for $i \in [n]$ to exact sampling on one of the five binary classification problems depicted in Table 3. To this purpose, we implement the rejection sampling (RS) detailed in Corollary 2 by using only one gradient descent step (i.e. we use the approximation $z_i^*(\theta) = \mathbf{A}_i \theta - \rho^2 \nabla U_i(\mathbf{A}_i \theta)$).

Figure 5 illustrates both the convergence rate towards high-probability regions between SGS with ULA and SGS with RS, and the acceptance probability associated to the RS scheme.
E.4 Comparison between SGS and SSGS

In this section, we compare SGS (Algorithm 1) with its stochastic version (Algorithm 2).

Figure 6 illustrates both the convergence rate towards high-probability regions and the classification scores w.r.t. the number of iterations $t$ for the SGS and SSGS with $\rho = 1$. The binary classification problem considered is associated to digits 3 and 5. For this problem, the number of samples is $n = 11,552$, and together with the prior distribution, we have $b = n + 1 = 11,553$ terms in the log-likelihood. SSGS was implemented by sampling $B$ terms without replacement, i.e. we have used the sets $B_j$ as all the $r = \binom{n}{B}$ subsets of size $B$ of the indices $1, \ldots, n$, plus the prior distribution (i.e. $B_j$ contains $B + 1$ terms). The weights $w_{ij}^{B_j}$ were chosen as 1 for the prior distribution ($i = b = n + 1$), and $b/B$ for all the others, and the probabilities were chosen uniformly as $p_i = 1/r$ for every $1 \leq i \leq r$. It is easy to check that the homogeneity Assumptions $(H_0) - (H_2)$ hold for these choices.
Figure 6: Logistic regression. Comparison between SGS and SSGS on the binary classification problem “5 versus 3” where $b = 11,553$. The parameter $\rho$ has been set to 1.

F Additional details on the image inpainting experiment in Section 5.2

F.1 Problem statement

We consider an image inpainting problem where an initial image $\theta \in \mathbb{R}^d$ (represented as a vector by lexicographic ordering) has to be retrieved from partial measurements $y \in \mathbb{R}^m$ ($m \ll d$ in general) under the Gaussian linear model

$$y = H\theta + \varepsilon,$$

(120)

where $H \in \mathbb{R}^{m \times d}$ is a decimation matrix associated to a damaging binary mask and $\varepsilon \sim \mathcal{N}(0_m, \sigma^2 I_m)$. Since the problem is ill-conditioned ($m < d$), we add some prior knowledge through the total variation (TV) prior distribution defined as

$$\pi_b(\theta) \propto \exp \left( -\tau \sum_{1 \leq i,j \leq \sqrt{d}} \| (\nabla \theta)_{i,j} \| \right),$$

where $\nabla \theta = (D^{(1)} \theta, D^{(2)} \theta)$ is the two-dimensional discrete gradient of $\theta$ and $\tau > 0$ is a regularization parameter. The operators $D^{(1)}, D^{(2)} \in \mathbb{R}^{d \times d}$ stand for the two first-order forward finite difference operators with appropriate boundary conditions on the vertical and horizontal directions, respectively.

Under this prior and the model (120), the Bayes’ rule leads to the non-differentiable posterior

$$\pi(\theta | y) \propto \exp \left( -\frac{1}{2\sigma^2} \| H\theta - y \|^2 - \tau \sum_{1 \leq i,j \leq \sqrt{d}} \| (\nabla \theta)_{i,j} \| \right).$$

F.2 Implementation details

Approximate instrumental model – Following the proposed approach, we set $b = 2$, $A_2 = \nabla$, $A_1 = H$ and introduce two auxiliary variables $z_b = (z_b^{(1)}, z_b^{(2)}) \in \mathbb{R}^{2d}$ and $z_1 \in \mathbb{R}^m$ such that the joint $\pi_\rho$ in (2) writes

$$\pi_\rho(\theta, z_{1:b} | y) \propto \exp \left( -\frac{1}{2\sigma^2} \| z_1 - y \|^2 - \tau \sum_{1 \leq i,j \leq \sqrt{d}} \| (z_b)_{i,j} \| \right) \times \exp \left( -\frac{1}{2\rho^2} \| z_1 - H\theta \|^2 - \frac{1}{2\rho^2} \| z_b - \nabla \theta \|^2 \right).$$
On color images – Since we considered color images, we only apply the MCMC algorithms to the illuminance channel. In this specific channel, the images have been sub-sampled and contaminated with Gaussian noise and all the performance criteria have been computed in this channel. Then, to display color images, we interpolated the color layers (Cb, Cr) by using bicubic interpolation. The parameter $\tau$ has been considered fixed and set to $\tau = 0.8$ for all the algorithms.

F.3 Sampling the auxiliary vector $z_1$

The conditional distribution associated to the auxiliary variable $z_1$ writes

$$
\pi_\rho(z_1|\theta, y) \propto \exp\left(-\frac{1}{2\sigma^2}||z_1 - y||^2 - \frac{1}{2\rho^2}||z_1 - H\theta||^2\right).
$$

(121)

This distribution is a Gaussian distribution $N(\mu_{z_1}, \Sigma_{z_1})$ where

$$
\Sigma_{z_1} = \frac{\rho^2\sigma^2}{\rho^2 + \sigma^2}I_n
$$

(122)

$$
\mu_{z_1} = \Sigma_{z_1}\left(\frac{H\theta}{\rho^2} + \frac{y}{\sigma^2}\right).
$$

(123)

F.4 Sampling the auxiliary vector $z_b$

The conditional distribution associated to the auxiliary variable $z_0$ writes

$$
\pi_\rho(z_b|\theta) \propto \exp\left(-\frac{1}{2\rho^2}||z_b - \nabla\theta||^2 - \frac{1}{2\rho^2}||z_0 - H\theta||^2\right).
$$

(124)

Thanks to the splitting of $\nabla\theta$, this conditional distribution can be sampled exactly by using data augmentation. Indeed, one can re-write the distribution involving the non-differentiable potential $||\cdot||$ as a mixture of normal and gamma distributions [32, Section 3.1]. Hence, sampling from (124) can be performed with the following two steps

1. Draw $\frac{1}{\gamma_k} \sim \text{InverseGaussian}\left(\frac{\tau}{||z_{b,k}||}, \tau^2\right) \forall k \in [d], \text{ if } ||z_{b,k}|| > 0$
2. Draw $\frac{1}{\gamma_k} \sim \text{InverseGaussian}\left(\frac{3}{2}, \frac{\tau^2}{2}\right) \forall k \in [d], \text{ if } ||z_{b,k}|| = 0$
3. Draw $z_{b,k}^{(1)} \sim N\left(\frac{\gamma_k(D(1)\theta)_k}{\rho^2 + \gamma_k}, \frac{\rho^2\gamma_k}{\rho^2 + \gamma_k}\right) \forall k \in [d],$
4. Draw $z_{b,k}^{(2)} \sim N\left(\frac{\gamma_k(D(2)\theta)_k}{\rho^2 + \gamma_k}, \frac{\rho^2\gamma_k}{\rho^2 + \gamma_k}\right) \forall k \in [d],$

where $z_{b,k}$ denotes the vector $(z_{b,k}^{(1)}, z_{b,k}^{(2)})$.

Note that all these sampling steps can be performed efficiently by “vectorizing” them.

F.5 Sampling the image of interest

The conditional distribution associated to the image to recover $\theta$ writes

$$
\pi_\rho(\theta|z_{1:b}) \propto \exp\left(-\frac{1}{2\rho^2}||z_b - \nabla\theta||^2 - \frac{1}{2\rho^2}||z_1 - H\theta||^2\right).
$$

(125)

The distribution [125] is a non-degenerate Gaussian distribution $N(\mu_\theta, \Sigma_\theta)$ where

$$
\Sigma_\theta = \rho^2 \left(\nabla^T \nabla + H^T H\right)^{-1}
$$

(126)
\[
\mu_\theta = \left( \nabla^T \nabla + H^T H \right)^{-1} \left( H^T z_1 + \nabla^T z_b \right). \tag{127}
\]

Note that \( \ker(H) \cap \ker(\nabla) = \{0_d\} \) which implies that the matrix \( M := H^T H + \nabla^T \nabla \) is non-singular. Sampling from this multivariate distribution can be done efficiently in \( O(d \log d) \) floating point operations by resorting to the two-dimensional discrete Fourier transform. Indeed, under periodic boundary conditions for \( \theta \), the matrix \( \nabla^T \nabla \) is a block circulant matrix and hence diagonalizable in the Fourier domain. For more details, we refer the interested reader to [57]. On the other hand, \( H^T H \) stands for a diagonal matrix with some zeros on the diagonal corresponding to the missing pixels. Since these two matrices cannot be diagonalized in the same domain, we use the auxiliary variable method of [37] to decouple them. Let \( \eta \|H^T H\|_S < \rho^2 \) where \( \|A\|_S \) is the spectral norm of the matrix \( A \). Then, we have the following two-step sampling scheme

\[
\text{Draw } v \sim \mathcal{N} \left( \frac{I_d}{\eta} - \frac{H^T H}{\rho^2} \right) \theta, \frac{I_d}{\eta} - \frac{H^T H}{\rho^2} \right) ,
\]

\text{Draw } \theta \sim \mathcal{N} \left( \mu_\theta, \Sigma_\theta \right),

where

\[
\Sigma_\theta = \left( \frac{I_d}{\eta} - \frac{\nabla^T \nabla}{\rho^2} \right)^{-1},
\]

\[
\mu_\theta = \Sigma_\theta \left( v + \frac{H^T}{\rho^2} z_1 + \frac{\nabla^T}{\rho^2} z_b \right).
\]

### F.6 Additional results

Figures 7 and 8 show the performance results of SGS with \( \rho = 1 \) on the A. Keys and A. Mauresmo images.

Tables 6 and 7 give performance results associated to the SGS for \( \rho = 0.5 \) and \( \rho = 10 \), respectively.
\[ \log \pi(\theta[t]) \]

Table 6: Image inpainting. Performances of the proposed algorithm (SGS, \( \rho = 0.5 \)) compared to the state-of-the-art MYULA for three images of celebrities. The performance criteria have been computed using the MMSE and averaged over 10 independent observations. \( 10^4 \) samples have been used for each algorithm and burn-in periods of 15000 and 12500 iterations have been considered for the SGS and MYULA, respectively.

| Image       | ISNR (dB)   | MSE          | time (s) |
|-------------|-------------|--------------|----------|
|             | MYULA       | SGS \( \rho = 0.5 \) | MYULA       | SGS \( \rho = 0.5 \) | MYULA       | SGS       |
| A. Keys     | 26.0 ± 0.3  | 26.1 ± 0.3   | 44.1 ± 3.0 | 43.3 ± 3.5   | 810         | 1030      |
| D. Beckham  | 26.3 ± 0.5  | 26.2 ± 0.5   | 31.6 ± 3.5 | 32.0 ± 3.9   | similar     | similar   |
| A. Mauresmo | 23.7 ± 0.3  | 23.7 ± 0.3   | 22.8 ± 1.6 | 22.7 ± 1.5   | similar     | similar   |

Table 7: Image inpainting. Performances of the proposed algorithm (SGS, \( \rho = 10 \)) compared to the state-of-the-art MYULA for three images of celebrities. The performance criteria have been computed using the MMSE and averaged over 10 independent observations. \( 10^4 \) samples have been used for each algorithm and burn-in periods of 100 and 12500 iterations have been considered for the SGS and MYULA, respectively.

| Image       | ISNR (dB)   | MSE          | time (s) |
|-------------|-------------|--------------|----------|
|             | MYULA       | SGS \( \rho = 10 \) | MYULA       | SGS \( \rho = 10 \) | MYULA       | SGS       |
| A. Keys     | 26.0 ± 0.3  | 16.3 ± 1.1   | 44.1 ± 3.0 | 425.9 ± 117.9 | 814         | 422       |
| D. Beckham  | 26.3 ± 0.5  | 15.5 ± 1.1   | 31.6 ± 3.5 | 387.6 ± 115.5 | similar     | similar   |
| A. Mauresmo | 23.7 ± 0.3  | 14.1 ± 1.7   | 22.8 ± 1.6 | 220.5 ± 107.4 | similar     | similar   |
References

[1] Baker, J., Fearnhead, P., Fox, E. B., and Nemeth, C. (2019). Control variates for stochastic gradient MCMC. *Statistics and Computing*, 29(3):599–615.

[2] Barbos, A.-C., Caron, F., Giovannelli, J.-F., and Doucet, A. (2017). Clone MCMC: parallel high-dimensional Gaussian Gibbs sampling. In *Advances in Neural Information Processing Systems*, pages 5020–5028.

[3] Bardenet, R., Doucet, A., and Holmes, C. (2014). Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach. In *Proceedings of the 31st International Conference on Machine Learning*.

[4] Bardenet, R., Doucet, A., and Holmes, C. (2017). On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18(47):1–43.

[5] Bennett, J. and Bez, N. (2015). Generating monotone quantities for the heat equation. *Journal für die reine und angewandte Mathematik (Crelles Journal)*.

[6] Bierkens, J., Fearnhead, P., and Roberts, G. (2019). The Zig-Zag process and super-efficient sampling for Bayesian analysis of big data. *The Annals of Statistics*, 47:1288–1320.

[7] Bouchard-Côté, A., Vollmer, S. J., and Doucet, A. (2018). The Bouncy Particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method. *Journal of the American Statistical Association*, 113:855–867.

[8] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122.

[9] Brosse, N., Durmus, A., and Moulines, E. (2018). The promises and pitfalls of stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems*, pages 8268–8278.

[10] Chambolle, A. (2004). An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20(1):89–97.

[11] Chambolle, A., Novaga, M., Cremers, D., and Pock, T. (2010). An introduction to total variation for image analysis. In *in Theoretical Foundations and Numerical Methods for Sparse Recovery*, De Gruyter.

[12] Chatterji, N. S., Flammarion, N., Ma, Y.-A., Bartlett, P. L., and Jordan, M. I. (2018). On the theory of variance reduction for stochastic gradient Monte Carlo. *arXiv preprint arXiv:1802.05431*.

[13] Chen, Z. and Vempala, S. S. (2019). Optimal convergence rate of Hamiltonian Monte Carlo for strongly logconcave distributions. *arXiv preprint arXiv:1905.02313*.

[14] Cheng, X. and Bartlett, P. (2018). Convergence of Langevin MCMC in KL-divergence. In Janoos, F., Mohri, M., and Sridharan, K., editors, *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 186–211. PMLR.

[15] Cheng, X., Chatterji, N. S., Bartlett, P. L., and Jordan, M. I. (2018). Underdamped Langevin MCMC: A non-asymptotic analysis. In Bubeck, S., Perchet, V., and Rigollet, P., editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 300–323. PMLR.

[16] Cornish, R., Vanetti, P., Bouchard-Côté, A., Deligiannidis, G., and Doucet, A. (2019). Scalable Metropolis–Hastings for exact Bayesian inference with large datasets. In *Proceedings of the 36th International Conference on Machine Learning*.

[17] Dai Pra, P., Scoppola, B., and Scoppola, E. (2012). Sampling from a Gibbs measure with pair interaction by means of PCA. *Journal of Statistical Physics*, 149(4):722–737.

[18] Dalalyan, A. S. and Karagulyan, A. (2019). User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*. 

46
[19] Dalalyan, A. S. and Riou-Durand, L. (2018). On sampling from a log-concave density using kinetic Langevin diffusions. arXiv preprint arXiv:1807.09382.

[20] Dubey, K. A., Reddi, S. J., Williamson, S. A., Poczos, B., Smola, A. J., and Xing, E. P. (2016). Variance reduction in stochastic gradient Langevin dynamics. In Advances in Neural Information Processing Systems, pages 1154–1162.

[21] Durmus, A., Majewski, S., and Miasojedow, B. (2019). Analysis of Langevin Monte Carlo via convex optimization. Journal of Machine Learning Research, 20(73):1–46.

[22] Durmus, A. and Moulines, E. (2017). Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. The Annals of Applied Probability, 27(3):1551–1587.

[23] Durmus, A. and Moulines, E. (2018). High-dimensional Bayesian inference via the unadjusted Langevin algorithm. arXiv preprint arXiv:1605.01559.

[24] Durmus, A., Moulines, E., and Pereyra, M. (2018). Efficient Bayesian computation by proximal Markov chain Monte Carlo: When Langevin meets Moreau. SIAM Journal on Imaging Sciences, 11(1):473–506.

[25] Dwivedi, R., Chen, Y., Wainwright, M. J., and Yu, B. (2018). Log-concave sampling: Metropolis-hastings algorithms are fast! In Conference On Learning Theory, pages 793–797.

[26] Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. Journal of the Royal Statistical Society. Series C (Applied Statistics), 41(2):337–348.

[27] Gradshteyn, I. S. and Ryzhik, I. M. (2015). Table of Integrals, Series, and Products. Academic Press, London, 8 edition.

[28] Hasenclever, L., Webb, S., Lienart, T., Vollmer, S., Lakshminarayanan, B., Blundell, C., and Teh, Y. W. (2017). Distributed Bayesian learning with stochastic natural gradient expectation propagation and the posterior server. Journal of Machine Learning Research, 18(1):3744–3780.

[29] Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. Journal of Machine Learning Research, 14(1):1303–1347.

[30] Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

[31] Korattikara, A., Chen, Y., and Welling, M. (2014). Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In Proceedings of the 31st International Conference on Machine Learning.

[32] Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. Bayesian Analysis, 5(2):369–411.

[33] Lawler, G. F. (2010). Random walk and the heat equation, volume 55. American Mathematical Society.

[34] Le Cun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324.

[35] Lemaréchal, C. and Sagastizábal, C. (1997). Practical aspects of the Moreau-Yosida regularization: Theoretical preliminaries. SIAM Journal on Optimization, 7(2):367–385.

[36] Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. Biometrika, 81(1):27–40.

[37] Marnissi, Y., Chouzenoux, E., Benazza-Benyahia, A., and Pesquet, J.-C. (2018). An auxiliary variable method for Markov chain Monte Carlo algorithms in high dimension. Entropy, 20(2).

[38] Minsker, S., Srivastava, S., Lin, L., and Dunson, D. (2014). Scalable and robust Bayesian inference via the median posterior. In Proceedings of the 31st International Conference on Machine Learning.
[39] Neiswanger, W., Wang, C., and Xing, E. P. (2014). Asymptotically exact, embarrassingly parallel MCMC. In Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence.

[40] Nesterov, Y. and Spokoiny, V. (2017). Random gradient-free minimization of convex functions. Foundations of Computational Mathematics, 17(2):527–566.

[41] Ollivier, Y. (2009). Ricci curvature of Markov chains on metric spaces. Journal of Functional Analysis, 256(3):810–864.

[42] Papandreou, G. and Yuille, A. L. (2010). Gaussian sampling by local perturbations. In Advances in Neural Information Processing Systems, pages 1858–1866.

[43] Park, T. and Casella, G. (2008). The Bayesian Lasso. Journal of the American Statistical Association, 103(482):681–686.

[44] Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2018). Speeding up MCMC by efficient data subsampling. Journal of the American Statistical Association. to appear.

[45] Rendell, L. J., Johansen, A. M., Lee, A., and Whiteley, N. (2018). Global consensus Monte Carlo. arXiv preprint arXiv:1807.09288.

[46] Robert, C. and Casella, G. (2013). Monte Carlo statistical methods. Springer Science & Business Media.

[47] Roberts, G. O. and Smith, A. F. (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. Stochastic processes and their applications, 49(2):207–216.

[48] Rockafellar, R. T. and Wets, R. J.-B. (1998). Variational analysis, volume 317 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin.

[49] Scheffé, H. (1947). A useful convergence theorem for probability distributions. The Annals of Mathematical Statistics, 18(3):434–438.

[50] Scott, S. (2017). Comparing consensus Monte Carlo strategies for distributed Bayesian computation. Brazilian Journal of Probability and Statistics.

[51] Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. (2016). Bayes and big data: The consensus Monte Carlo algorithm. International Journal of Management Science and Engineering Management, 11:78–88.

[52] Villani, C. (2008). Optimal Transport: Old and New. Springer Berlin Heidelberg.

[53] Vono, M., Dobigeon, N., and Chainais, P. (2019a). Asymptotically exact data augmentation: models, properties and algorithms. arXiv preprint arXiv:1902.05754.

[54] Vono, M., Dobigeon, N., and Chainais, P. (2019b). Split-and-augmented Gibbs sampler - Application to large-scale inference problems. IEEE Transactions on Signal Processing, 67(6):1648–1661.

[55] Wang, X. and Dunson, D. B. (2013). Parallelizing MCMC via Weierstrass sampler. arXiv preprint arXiv:1312.4605.

[56] Wang, X., Guo, F., Heller, K. A., and Dunson, D. B. (2015). Parallelizing MCMC with random partition trees. In Advances in Neural Information Processing Systems.

[57] Wang, Y., Yang, J., Yin, W., and Zhang, Y. (2008). A new alternating minimization algorithm for total variation image reconstruction. SIAM Journal on Imaging Sciences, 1(3):248–272.

[58] Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In Proceedings of the 28th International Conference on Machine Learning, pages 681–688.