On free energy barriers in Gaussian priors and failure of cold start MCMC for high-dimensional unimodal distributions

Originally published in:
Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 381(2247), https://doi.org/10.1098/rsta.2022.0150

PHILOSOPHICAL TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

Research

Cite this article: Bandeira AS, Maillard A, Nickl R, Wang S. 2023 On free energy barriers in Gaussian priors and failure of cold start MCMC for high-dimensional unimodal distributions. *Phil. Trans. R. Soc. A* **381**: 20220150. https://doi.org/10.1098/rsta.2022.0150

Received: 5 September 2022
Accepted: 17 November 2022

One contribution of 16 to a theme issue 'Bayesian inference: challenges, perspectives, and prospects'.

Subject Areas:
statistics

Keywords:
MCMC, Bayesian inference, Gaussian processes, computational hardness

Author for correspondence:
Richard Nickl
e-mail: nickl@maths.cam.ac.uk

On free energy barriers in Gaussian priors and failure of cold start MCMC for high-dimensional unimodal distributions

Afonso S. Bandeira[1], Antoine Maillard[1], Richard Nickl[2] and Sven Wang[3]

[1]Department of Mathematics, ETH Zürich, Zurich, Switzerland
[2]Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge, UK
[3]Institute for Data, Systems and Society, MIT, Cambridge, MA, USA

RN, 0000-0002-0889-5422

We exhibit examples of high-dimensional unimodal posterior distributions arising in nonlinear regression models with Gaussian process priors for which Markov chain Monte Carlo (MCMC) methods can take an exponential run-time to enter the regions where the bulk of the posterior measure concentrates. Our results apply to worst-case initialized ('cold start') algorithms that are local in the sense that their step sizes cannot be too large on average. The counter-examples hold for general MCMC schemes based on gradient or random walk steps, and the theory is illustrated for Metropolis–Hastings adjusted methods such as preconditioned Crank–Nicolson and Metropolis-adjusted Langevin algorithm.

This article is part of the theme issue 'Bayesian inference: challenges, perspectives, and prospects'.

1. Introduction

Markov chain Monte Carlo (MCMC) methods are the workhorse of Bayesian computation when closed formulae for estimators or probability distributions are not available. For this reason they have been central to the development and success of high-dimensional

THE ROYAL SOCIETY PUBLISHING

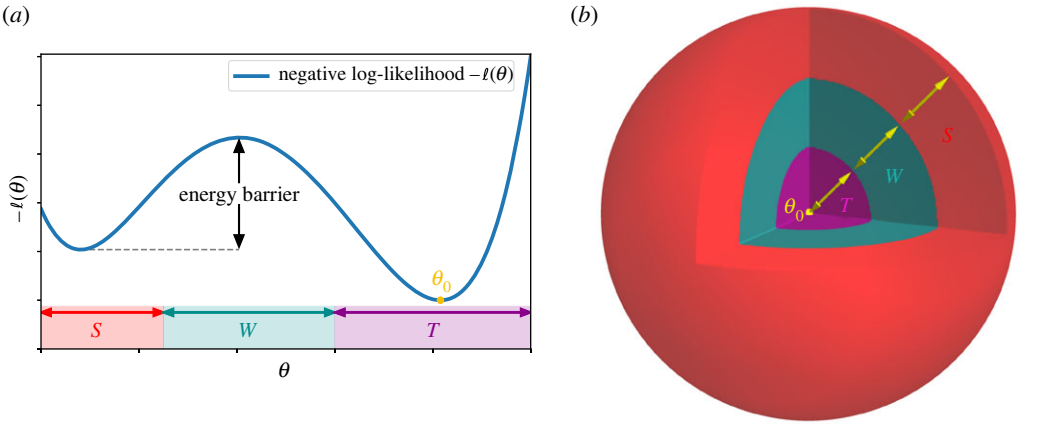


Figure 1. Two possible sources of MCMC hardness in high dimensions: multi-modal likelihoods and entropic barriers. (*a*) In low dimensions (here $D = 1$), MCMC hardness usually arises because of a non-unimodal likelihood, creating an 'energy barrier', even though the maximum likelihood is attained at $\theta = \theta_0$. The MCMC algorithm is assumed to be initialized in the set $\mathcal{S}$ containing a local maximum of the likelihood. (*b*) Illustration of the arising of entropic (or volumetric) difficulties, here in dimension $D = 3$: the set of points close to θ_0 has much less volume than the set of points far away. As D increases, this phenomenon is amplified: all ratios of volumes of the three sets $\mathcal{T}, \mathcal{W}, \mathcal{S}$ scale exponentially with D. (Online version in colour.)

Bayesian statistics in the last decades, where one attempts to generate samples from some *posterior distribution* $\Pi(\cdot|\text{data})$ arising from a prior Π on D-dimensional Euclidean space and the observed data vector. MCMC methods tend to perform well in a large variety of problems, are very flexible and user-friendly, and enjoy many theoretical guarantees. Under mild assumptions, they are known to converge to their stationary 'target' distributions as a consequence of the ergodic theorem, albeit perhaps at a slow speed, requiring a large number of iterations to provide numerically accurate algorithms. When the target distribution is log-concave, MCMC algorithms are known to mix rapidly, even in high dimensions. But for general D-dimensional densities, we have only a restricted understanding of the scaling of the mixing time of Markov chains with D or with the 'informativeness' (sample size or noise level) of the data vector.

A classical source of difficulty for MCMC algorithms are multi-modal distributions. When there is a deep well in the posterior density between the starting point of an MCMC algorithm and the location where the posterior is concentrated, many MCMC algorithms are known to take an exponential time—proportional to the depth of the well—when attempting to reach the target region, even in low-dimensional settings, see figure 1*a* and also the discussion surrounding proposition 4.2 below. However, for distributions with a single mode and when the dimension D is fixed, MCMC methods can usually be expected to perform well.

In essence this article is an attempt to explain how, in high dimensions, wells can be formed *without* multi-modality of a given posterior distribution. The difficulty in this case is volumetric, also referred to as *entropic*: while the target region contains most of the posterior mass, its (prior) volume is so small compared to the rest of the space that an MCMC algorithm may take an exponential time to find it, see figure 1*b*. This competition between 'energy'—here represented by the log-likelihood ℓ_N in the posterior distribution $\mathrm{d}\Pi(\cdot|\text{data}) = \exp\{\ell_N + \log \mathrm{d}\pi\}$—and 'entropy' (related to the prior term π) has also been exploited in recent work on statistical aspects of MCMC in various high dimensional inference and statistical physics models [1–5]. These ideas somewhat date back to the nineteenth century foundations of statistical mechanics [6] and the notion of free energy, consisting of a sum of energetic and entropic contributions which the system spontaneously attempts to minimize. The 'MCMC-hardness' phenomenon described above is then akin to the meta-stable behaviour of thermodynamical systems, such as glasses or supercooled liquids. As the temperature decreases, such systems can undergo a 'first-order' phase transition, in which a global free energy minimum (analoguous to the

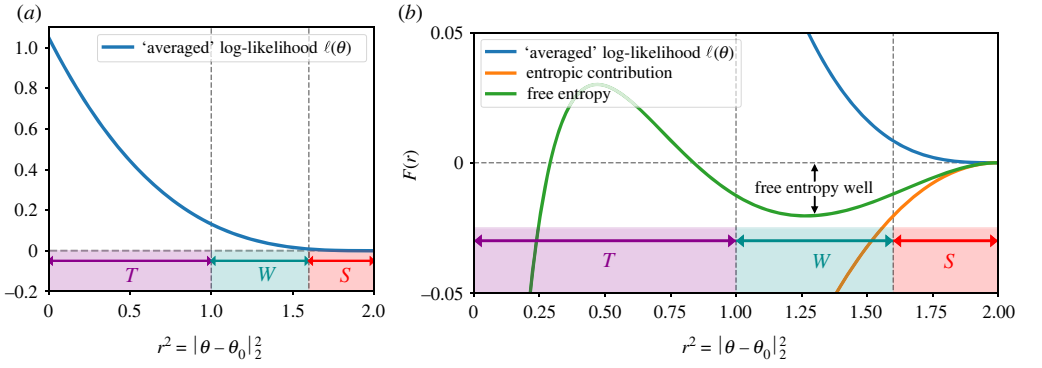


Figure 2. Illustration of a free-energy barrier (or free-entropy well) arising with a unimodal posterior. The model is an 'averaged' version of the spiked tensor model, with log-likelihood $\ell_n(\theta) = \lambda\langle\theta, \theta_0\rangle^3/2$ and uniform prior Π on the n-dimensional unit sphere $\mathbb{S}^{n-1}$. θ_0 is chosen arbitrarily on $\mathbb{S}^{n-1}$. The posterior is $\mathrm{d}\Pi(\theta|Y) \propto \exp\{n\ell_n(\theta)\}\,\mathrm{d}\Pi(\theta)$, for $\theta \in \mathbb{S}^{n-1}$. Up to a constant, the free entropy $F(r) = (1/n)\log\int \mathrm{d}\Pi(\theta|Y)\delta(r - ||\theta - \theta_0||_2)$ can be decomposed as the sum of $\ell_n(\theta)$ (that only depends on $r = ||\theta - \theta_0||_2$) and the 'entropic' contribution $(1/n)\log\int \mathrm{d}\Pi(\theta)\delta(r - ||\theta - \theta_0||_2)$. In the figure we show $\lambda = 2.1$. (Online version in colour.)

target region above) abruptly appears, while the system remains trapped in a suboptimal local minimum of the free energy (the starting region of the MCMC algorithm). For the system to go to thermodynamic equilibrium it must cross an extensive free energy barrier: such a crossing requires an exponentially long time, so that the system appears equilibrated on all relevant timescales, similarly to the MCMC stuck in the starting region. Classical examples include glasses and the popular experiment of rapid freezing of supercooled water (i.e. water that remained liquid at negative temperatures) after introducing a perturbation.

Inspired by recent work [4,5,7], let us illustrate some of the volumetric phenomena which are key to our results below. We separate the parameter space into three regions (see figures 1 and 2), which we name by common MCMC terminology. Firstly a *starting* (or initialization) region $\mathcal{S}$, where an algorithm starts, secondly a *target* region $\mathcal{T}$ where both the bulk of the posterior mass and the ground truth are situated, and thirdly an intermediate *free-entropy well*[1] $\mathcal{W}$ that separates $\mathcal{S}$ from $\mathcal{T}$.[2] In our theorems, these regions will be characterized by their Euclidean distance to the ground truth parameter θ_0 generating the data. The prior volumes of the ϵ-annuli $\{\theta : r - \epsilon < ||\theta - \theta_0||_2 \le r\}, r > 0$, closer to the ground truth are smaller than those further out as illustrated in figure 1*b*, and in high dimensions this effect becomes quantitative in an essential way. Specifically, the trade-off between the entropic and energetic terms can happen such that the following three statements are simultaneously true.

(i) $\mathcal{T}$ contains 'almost all' of the posterior mass.
(ii) As one gets closer to $\mathcal{T}$ (and thus the ground truth θ_0), the log-likelihood is strictly monotonically increasing.
(iii) Yet $\mathcal{S}$ still possesses exponentially more posterior mass than $\mathcal{W}$.

Using 'bottleneck' arguments from Markov chain theory (ch. 7 in [8]), this means that an MCMC algorithm that starts in $\mathcal{S}$ is expected to take an exponential time to visit $\mathcal{W}$. If the step size is such that it cannot 'jump over' $\mathcal{W}$, this also implies an exponential hitting time lower bound for reaching $\mathcal{T}$. This is illustrated in figure 2 for an averaged version of the model described in §2.

In the situation described above, the MCMC iterates never visit the region where the posterior is statistically informative, and hence yield no better inference than a random number generator.

[1]As classical in statistical physics, we call free entropy the negative of the free energy.

[2]In a physical system, these regions would correspond respectively to a region including a meta-stable state, a region including the globally stable state and a free energy barrier.

One could regard this as a 'hardness' result about computation of posterior distributions in high dimensions by MCMC. In this work we show that such situations can occur generically and establish hitting time lower bounds for common gradient or random walk based MCMC schemes in model problems with nonlinear regression and Gaussian process priors. Before doing this, we briefly review some important results of Ben Arous *et al.* [4] for the problem of principle component analysis (PCA) in tensor models, from which the inspiration for our work was drawn. This technique to establish lower bounds for MCMC algorithms has also recently been leveraged in [5] in the context of sparse PCA, and in [7] to establish connections between MCMC lower bounds and the Low Degree Method for algorithmic hardness predictions (see [9] for an expository note on this technique).

When the target distribution is globally log-concave, pictures such as in figure 2 are ruled out (see also remark 4.7) and polynomial-time mixing bounds have been shown for a variety of commonly used MCMC methods. While an exhaustive discussion would be beyond the scope of this paper, we mention here the seminal works [10,11] which were among the first to demonstrate high-dimensional mixing of discretized Langevin methods (even upon 'cold-start' initializations like the ones assumed in the present paper). In concrete nonlinear regression models, polynomial-time computation guarantees were given in [12] under a general 'gradient stability' condition on the regression map which guarantees that the posterior is (with high probability) locally log-concave on a large enough region including θ_0. While this condition can be expected to hold under natural injectivity hypotheses and was verified for an inverse problem with the Schrödinger equation in [12], for non-Abelian X-ray transforms in [13], the 'Darcy flow' model involving elliptic partial differential equations (PDE) in [14] and for generalized linear models in [15], all these results hinge on the existence of a suitable initializer of the gradient MCMC scheme used. These results form part of a larger research programme [14,16–19] on algorithmic and statistical guarantees for Bayesian inversion methods [20] applied to problems with partial differential equations. The present article shows that the hypothesis of existence of a suitable initializer is—at least in principle—essential in these results if $D/N \to \kappa > 0$, and that at most 'moderately' high-dimensional ($D = \mathcal{O}(N)$) MCMC implementations of Gaussian process priors may be preferable to bypass computational bottlenecks.

Our negative results apply to (worst-case initialized) Markov chains whose step sizes cannot be too large with high probability. As we show this includes many commonly used algorithms (such as preconditioned Crank–Nicolson (pCN) and Metropolis-adjusted Langevin algorithm (MALA)) whose dynamics are of a 'local' nature. There are a variety of MCMC methods developed recently, such as piece-wise deterministic Markov processes, boomerang or zig-zag samplers [21–24] which may not fall into our framework. While we are not aware of any rigorous results that would establish polynomial hitting or mixing times of these algorithms for high-dimensional posterior distributions such as those exhibited here, it is of great interest to study whether our computational hardness barriers can be overcome by 'non-local' methods. There is some empirical evidence that this may be possible. For instance, in the numerical simulation of models of supercooled liquids [25], methods such as swap Monte Carlo [26] have been observed to equilibrate to low-temperature distributions which were not reachable by local approaches. Another example is given by the planted clique problem [27]: this model is conjectured to possess a large algorithmically hard phase, and local Monte Carlo methods are known to fail far from the conjectured algorithmic threshold [28–30]. On the other hand, non-local exchange Monte Carlo methods (such as parallel tempering [31]) have been numerically observed to perform significantly better [32].

2. The spiked tensor model: an illustrative example

In this section, we present (a simplified version of) results obtained mostly in [4]. First some notation. For any $n \geq 1$, we denote by $\mathbb{S}^{n-1} = \{\theta \in \mathbb{R}^n : ||\theta||_2 = 1\}$ the Euclidean unit sphere in n dimensions. For $\theta, \theta' \in \mathbb{R}^n$ we denote $\theta \otimes \theta' = (\theta_i \theta'_j)_{1 \leq i,j \leq n} \in \mathbb{R}^{n^2}$ their tensor product.

Spiked tensor estimation is a synthetic model to study tensor PCA, and corresponds to a Gaussian Additive Model with a low-rank prior. More formally, it can be defined as follows [33].

Definition 2.1 (Spiked tensor model). Let $p \geq 3$ denote the order of the tensor. The observations Y and the parameter θ are generated according to the following joint probability distribution:

$$\mathrm{d}\mathbb{Q}(Y,\theta) = \frac{1}{(2\pi)^{n^p/2}} \exp\left\{-\frac{1}{2}\left|\left|Y - \sqrt{n}\lambda\theta^{\otimes p}\right|\right|_2^2\right\} \mathrm{d}\Pi(\theta)\,\mathrm{d}Y. \tag{2.1}$$

Here, $\mathrm{d}Y$ denotes the Lebesgue measure on the space $(\mathbb{R}^n)^{\otimes p} = \mathbb{R}^{n^p}$ of p-tensors of size n. Π is the uniform probability measure on $\mathbb{S}^{n-1}$, and $\lambda \geq 0$ is the signal-to-noise ratio (SNR) parameter. In particular, the posterior distribution $\Pi(\theta|Y)$ is

$$\mathrm{d}\Pi(\theta|Y) = \frac{1}{\mathcal{Z}_Y} \exp(\ell_{n,Y}(\theta))\,\mathrm{d}\Pi(\theta), \tag{2.2}$$

in which $\mathcal{Z}_Y$ is a normalization, and we defined the *log-likelihood* (up to additive constants) as

$$\ell_{n,Y}(\theta) = \frac{1}{2}\sqrt{n}\lambda\langle\theta^{\otimes p}, Y\rangle. \tag{2.3}$$

In the following, we study the model from definition 2.1 via the prism of statistical inference. In particular, we will study the posterior $\Pi(\theta|Y)$ for a fixed[3] 'data tensor' Y. Since such a tensor was generated according to the marginal of (2.1), we parameterize it as $Y = \lambda\sqrt{n}\theta_0^{\otimes p} + Z$, with Z a p-tensor with i.i.d. $\mathcal{N}(0,1)$ coordinates, and θ_0 a 'ground truth' vector uniformly sampled in $\mathbb{S}^{n-1}$. The goal of our inference task is to recover information on the low-rank perturbation $\theta_0^{\otimes p}$ (or equivalently on the vector θ_0, possibly up to a global sign depending on the parity of p) from the posterior distribution $\Pi(\cdot|Y)$.

Crucially, we are interested in the limit of the model of definition 2.1 as $n \to \infty$. In particular, all our statements, although sometimes non-asymptotic, are to be interpreted as n grows. We say that an event occurs 'with high probability' (w.h.p.) when its probability is $1 - \mathcal{O}_n(1)$.[4] Moreover, by rotation invariance, all statements are uniform over $\theta_0 \in \mathbb{S}^{n-1}$, so that said probabilities only refer to the noise tensor Z. Finally, throughout our discussion we will work with latitude intervals (or bands) on the sphere, with the North Pole taken to be θ_0. We characterize them using inner products (correlations) $\langle\theta,\theta_0\rangle$ for odd p, and $|\langle\theta,\theta_0\rangle|$ for even p (since in this case θ_0 and $-\theta_0$ are indistinguishable from the point of view of the observer).

Definition 2.2 (Latitude intervals). Assume that $p \geq 3$ is even. For $0 \leq s < t \leq 1$ we define:

— $\mathcal{S}_s = \{\theta \in \mathbb{S}^{n-1} : |\langle\theta,\theta_0\rangle| \leq s\}$,
— $\mathcal{W}_{s,t} = \{\theta \in \mathbb{S}^{n-1} : s < |\langle\theta,\theta_0\rangle| \leq t\}$,
— $\mathcal{T}_t = \{\theta \in \mathbb{S}^{n-1} : t < |\langle\theta,\theta_0\rangle|\}$.

If p is odd, we define these sets similarly, replacing $|\langle\theta,\theta_0\rangle|$ by $\langle\theta,\theta_0\rangle$.

Note that these sets can also be characterized using the distance to the ground truth, e.g. $\mathcal{S}_s = \{\theta \in \mathbb{S}^{n-1} : \min\{||\theta - \theta_0||_2^2, ||\theta + \theta_0||_2^2\}| \geq 2(1-s)\}$ when p is even.

(a) Posterior contraction

We can use uniform concentration of the likelihood to show that as $\lambda \to \infty$ (*after* taking the limit $n \to \infty$) the posterior contracts in a region infinitesimally close to the ground truth θ_0. We first show that a region arbitrarily close to the ground truth exponentially dominates a very large starting region:

[3]Note that we assume here that the statistician has access to the distribution $\Pi(\cdot|Y)$ (and in particular to λ), a setting sometimes called *Bayes-optimal* in the literature.

[4]Often the $\mathcal{O}_n(1)$ term will be exponentially small, but we will not require such a strong control.

Proposition 2.3. *For any $K > 0$ there exists $\lambda_0 > 0$ and functions $\{s(\lambda), t(\lambda)\} \in [0,1)$ such that $s(\lambda) < t(\lambda)$, $\{s(\lambda), t(\lambda)\} \to 1$ as $\lambda \to \infty$ and for all $\lambda \geq \lambda_0$:*

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{\Pi(\mathcal{S}_{s(\lambda)}|Y)}{\Pi(\mathcal{T}_{t(\lambda)}|Y)} \leq -K, \quad \textit{almost surely.} \tag{2.4}$$

Posterior contraction is the content of the following result:

Corollary 2.4 (Posterior contraction). *There exists $\lambda_0 > 0$ and a function $s(\lambda) \in [0,1)$ satisfying $s(\lambda) \to 1$ as $\lambda \to \infty$, such that for all $\lambda \geq \lambda_0$:*

$$\lim_{n\to\infty} \Pi[\mathcal{T}_{s(\lambda)}|Y] = 1, \quad \textit{almost surely.} \tag{2.5}$$

The proofs of proposition 2.3 and corollary 2.4 are given in appendix A.

Remark 2.5 (Suboptimality of uniform bounds). Stronger than corollary 2.4, it is known that there exists a sharp threshold $\lambda^\star(p)$ such that for any $\lambda > \lambda^\star(p)$ the posterior mean, as well as the maximum likelihood estimator, sit w.h.p. in $\mathcal{T}_{s(\lambda)}$, with $s(\lambda) > 0$, while such a statement is false for $\lambda \leq \lambda^\star(p)$ [34–36]. The λ_0 given by corollary 2.4 is, on the other hand, clearly not sharp, because of the crude uniform bound used in the proof. This can easily be understood in the $p = 2$ case, corresponding to rank-one matrix estimation: uniform bounds such as the ones used here would show posterior contraction for $\lambda = \omega(1)$, while it is known through the celebrated BBP transition that the maximum likelihood estimator is already correlated with the signal for any $\lambda > 1$ [37]. With more refined techniques from the study of random matrices and spin glass theory of statistical physics it is often possible to obtain precise constants for such relevant thresholds.

(b) Algorithmic bottleneck for MCMC

Simple volume arguments, associated with an ingenious use of Markov's inequality due to Ben Arous *et al.* [4] and of the rotation invariance of the noise tensor Z, allow us to get a computational hardness result for MCMC algorithms, even though the posterior contracts infinitesimally close to the ground truth as we saw in corollary 2.4. In the context of the spiked tensor model, these computational hardness results can be found in [4] (see in particular §7). We will state similar results for general nonlinear regression models in §3: in this context we will not need to use the Markov's inequality-based technique of Ben Arous *et al.* [4], and will solely rely on concentration arguments.

Recall that by §2(a), we can find $s(\lambda)$ such that $s(\lambda) \to 1$ as $\lambda \to \infty$ and for all λ large enough $\Pi(\mathcal{T}_{s(\lambda)}|Y) = 1 - \mathcal{O}_n(1)$. Here, we show that escaping the 'initialization' region of the MCMC algorithm is hard in a large range of λ (possibly diverging with n). In what follows, the step size of the algorithm denotes the maximal change $||x^{t+1} - x^t||_2$ allowed in any iteration.[5] We first state this bottleneck result informally.

Proposition 2.6 (MCMC bottleneck, informal). *Assume that $\lambda = \mathcal{O}(n^{(p-2)/4+\eta})$ for all $\eta > 0$. Then any MCMC algorithm whose invariant distribution is $\Pi(\cdot|Y)$, and with a step size bounded by $\delta = \mathcal{O}([n\lambda^2]^{-1/p})$, will take an exponential time to get out of the 'initialization' region.*

Note that the step size condition of proposition 2.6 is always meaningful, since our hypothesis on λ implies $[n\lambda^2]^{-1/p} = \omega(n^{-1/2})$, and many MCMC algorithms (e.g. any procedure in which a number $\mathcal{O}(1)$ of coordinates of the current iterate are changed in a single iteration) will have a step size $\mathcal{O}(n^{-1/2})$.

Remark 2.7. The results of Ben Arous *et al.* [4] are stated when considering for the invariant distribution of the MCMC a more general 'Gibbs-type' distribution $\mathbb{G}_{\beta,Y}(\mathrm{d}x) \propto \mathrm{e}^{\beta H(x)}\,\mathrm{d}\Pi(x)$, with $H(x) = (\sqrt{n}/2)\langle x^{\otimes p}, Y\rangle$. The case we consider here is the 'Bayes-optimal' $\beta = \lambda$, for which $\mathbb{G}_{\lambda,Y} = \Pi(\cdot|Y)$. For the general distribution $\mathbb{G}_{\beta,Y}$ the conditions of proposition 2.6 become

[5] As we will detail in the following sections, see assumption 3.1, the statements remain true if the change is allowed to be higher than the required maximum with exponentially small probability.

$\beta\lambda = \mathcal{O}(n^{(p-2)/2+\eta})$ and $\delta = \mathcal{O}[(n\beta\lambda)^{-1/p}]$. The authors of Ben Arous *et al.* [4] usually consider $\beta = \mathcal{O}(1)$, so that they show the bottleneck under the condition $\lambda = o(n^{(p-2)/2+\eta})$.

More generally, $\lambda \ll n^{(p-2)/4}$ is conjectured to be a regime in which *all* polynomial-time algorithms fail to recover θ_0 [33,38–41]. On the other hand, 'local' methods (such as gradient-based algorithms [42–46], message-passing iterations [35] or natural MCMC algorithms such as the ones of previous remark) are conjectured or known to fail in the larger range $\lambda \ll n^{(p-2)/2}$. Proposition 2.6 shows that 'Bayes-optimal' MCMC algorithms fail for $\lambda \ll n^{(p-2)/4}$. To the best of our knowledge, analysing this class of algorithms in the regime $n^{(p-2)/4} \ll \lambda \ll n^{(p-2)/2}$ is still open.

Let us now state formally the key ingredient behind proposition 2.6. It is a rewriting of the 'free energy wells' result of Ben Arous *et al.* [4].

Lemma 2.8 (Bottleneck, formal). *Assume that* $\lambda = \mathcal{O}(n^{(p-2)/4+\eta})$ *for all* $\eta > 0$, *and let* $\delta = \mathcal{O}([n\lambda^2]^{-1/p})$. *Let* $r(\varepsilon) = n^{-1/2+\varepsilon}$. *Then for any* $\varepsilon > 0$ *small enough, there exists* $c, C > 0$ *such that for large enough n, with probability at least* $1 - \exp(-cn^{2\varepsilon})$ *we have:*

$$\frac{\Pi(\mathcal{S}_{r(\varepsilon)}|Y)}{\Pi(\mathcal{W}_{r(\varepsilon),r(\varepsilon)+\delta}|Y)} \geq \exp\{Cn^{2\varepsilon}\}. \tag{2.6}$$

Note that by simple volume arguments, $\Pi(\mathcal{S}_{r(\varepsilon)}) = 1 - o_n(1)$, so that $\mathcal{S}_{r(\varepsilon)}$ contains 'almost all' the mass of the uniform distribution.

One can then deduce from lemma 2.8 hitting time lower bounds for MCMCs using a folklore bottleneck argument—see Jerrum [8]—that we recall here in a simplified form (see also [5], as well as proposition 4.4, where we will detail it further along with a short proof).

Proposition 2.9. *We fix any Y and n, and let any* $0 < s < t < 1$. *Let* $\theta^{(0)}, \theta^{(1)}, \ldots$ *be a Markov chain on* $\mathbb{S}^{n-1}$ *with stationary distribution* $\Pi(\cdot|Y)$, *and initialized from* $\theta^{(0)} \sim \Pi_{\mathcal{S}_s}(\cdot|Y)$, *the posterior distribution conditioned on* $\mathcal{S}_s$. *Let* $\tau_t = \inf\{k \in \mathbb{N} : \theta^{(k)} \in \mathcal{T}_t\}$ *be the hitting time of the Markov chain onto* $\mathcal{T}_t$. *Then, for any* $k \geq 1$,

$$\Pr(\tau_t \leq k) \leq k\frac{\Pi(\mathcal{W}_{s,t}|Y)}{\Pi(\mathcal{S}_s|Y)}. \tag{2.7}$$

Remark 2.10 (MCMC initialization). Note that lemma 2.8, combined with proposition 2.9, shows hardness of MCMC initialized in points drawn from $\Pi_{\mathcal{S}_{r(\epsilon)}}(\cdot|Y)$. In particular, it is easy to see that this implies (via the probabilistic method) the existence of such 'hard' initializing points. While one might hope to show such negative results for more general initialization, this remains an open problem. On the other hand, Ben Arous *et al.* [4] shows that there exists initializers in $\mathcal{S}_{r(\epsilon)}$ for which vanilla Langevin dynamics achieve non-trivial recovery of the signal even for $\lambda = \Theta_n(1)$ (a phenomenon they call 'equatorial passes').

3. Main results for nonlinear regression with Gaussian priors

We now turn to the main contribution of this article, which is to exhibit some of the phenomena described in §2 in the context of nonlinear regression models. All the theorems of this section are proven in detail in §4.

Consider data $Z^{(N)} =^{\text{iid}} (Y_i, X_i)_{i=1}^N$ from the random design regression model

$$Y_i = \mathcal{G}(\theta)(X_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0,1),\ i = 1, \ldots, N, \tag{3.1}$$

where $\mathcal{G} : \Theta \to L^2_\mu(\mathcal{X})$ is a regression map taking values in the space $L^2(\mathcal{X}) = L^2_\mu(\mathcal{X})$ on some bounded subset $\mathcal{X}$ of $\mathbb{R}^d$, and where the $X_i \sim^{\text{iid}} \mu$ are drawn uniformly on $\mathcal{X}$. For convenience, we assume that $\mathcal{X}$ has Lebesgue measure $\int_{\mathcal{X}} \mathrm{d}x = 1$. The law of the data $\mathrm{d}P_\theta^N(z_1, \ldots, z_N)$

$= \prod_{i=1}^{N} \mathrm{d}P_\theta(z_i)$ is a product measure on $(\mathbb{R} \times \mathcal{X})^N$, with associated expectation operator E_θ^N. Here θ varies in some parameter space

$$\Theta \subseteq \mathbb{R}^D, \quad \frac{D}{N} \simeq \kappa \geq 0,$$

and $\theta_0 \in \Theta$ is a 'ground truth' (we could use 'mis-specified' θ_0 and project it onto Θ). We will primarily consider the case where $\kappa > 0$ and $\Theta = \mathbb{R}^D$, and consider high-dimensional asymptotics where D (and then also N) diverge to infinity, even though some aspects of our proofs do not rely on these assumptions. We will say that events A_N hold with high probability if $P_{\theta_0}^N(A_N) \to 1$ as $N \to \infty$, and we will use the same terminology later when it involves the law of some Markov chain.

Let Π be a prior (Borel probability measure) on Θ so that given the data $Z^{(N)}$ the posterior measure is the 'Gibbs'-type distribution

$$\mathrm{d}\Pi(\theta|Z^{(N)}) = \frac{\mathrm{e}^{\ell_N(\theta)}\,\mathrm{d}\Pi(\theta)}{\int_\Theta \mathrm{e}^{\ell_N(\theta)}\,\mathrm{d}\Pi(\theta)}, \quad \theta \in \Theta, \tag{3.2}$$

where

$$\ell_N(\theta) = -\frac{1}{2}\sum_{i=1}^{N} |Y_i - \mathcal{G}(\theta)(X_i)|^2, \quad \ell(\theta) = \mathrm{E}_{\theta_0}^N \ell_N(\theta), \ \ \theta \in \Theta.$$

(a) Hardness examples for posterior computation with Gaussian priors

We are concerned here with the question of whether one can sample from the Gibbs' measure (3.2) by MCMC algorithms. The priors will be Gaussian, so the 'source' of the difficulty will arise from the log-likelihood function ℓ_N. On the one hand, recent work [10–14] has demonstrated that if $\ell_N(\theta)$ is 'on average' (under E_{θ_0}) log-concave, possibly only just locally near the ground truth θ_0, then MCMC methods that are initialized into the area of log-concavity can mix towards $\Pi(\cdot|Z^{(N)})$ in polynomial time even in high-dimensional ($D \to \infty$) and 'informative' ($N \to \infty$) settings. In the absence of such structural assumptions, however, posterior computation may be intractable, and the purpose of this section is to give some concrete examples for this with choices of $\mathcal{G}$ that are representative for nonlinear regression models.

We will provide lower bounds on the run-time of 'worst case' initialized MCMC in settings where the average posterior surface is not *globally* log-concave but still unimodal. Both the log-likelihood function and posterior density exhibit linear growth towards their modes, and the average log-likelihood is locally log-concave at θ_0. In particular, the Fisher information is well defined and non-singular at the ground truth.

The computational hardness does not arise from a local optimum ('multi-modality'), but from the difficulty MCMC encounters in 'choosing' among many high-dimensional directions when started away from the bulk of the support of the posterior measure. That such problems occur in high dimensions is related to the probabilistic structure of the prior Π, and the manifestation of 'free energy barriers' in the posterior distribution.

In many applications of Bayesian statistics, such as in machine learning or in nonlinear inverse problems with PDEs, *Gaussian process priors* are commonly used for inference. To connect to such situations we illustrate the key ideas that follow with two canonical examples where the prior on $\mathbb{R}^D$ is the law

$$(a)\ \theta \sim \mathcal{N}\left(0, \frac{I_D}{D}\right), \quad \text{or } (b)\ \theta \sim \mathcal{N}(0, \Sigma_\alpha), \tag{3.3}$$

where Σ_α is the covariance matrix arising from the law of a D-dimensional Whittle–Matérn-type Gaussian random field (see §4(d)(i) for a detailed definition). These priors represent widely popular choices in Bayesian statistical inference [47,48] and can be expected to yield consistent statistical solutions of regression problems even when $D/N \geq \kappa > 0$, see [48,49]. In (b), we can also accommodate a further 'rescaling' (N-dependent shrinkage) of the prior similar to what has been used in recent theory for nonlinear inverse problems [12,13,18], see remark 4.6 for details.

We will present our main results for the case where the ground truth is $\theta_0 = 0$. This streamlines notation while also being the 'hardest' case for negative results, since the priors from (a) and (b) are then already centred at the correct parameter.

To formalize our results, let us define balls

$$B_r = \{\theta \in \mathbb{R}^D : ||\theta||_{\mathbb{R}^D} \leq r\}, \quad r > 0, \tag{3.4}$$

centred at $\theta_0 = 0$. We will also require the annuli

$$\Theta_{r,\varepsilon} = \{\theta \in \mathbb{R}^D : ||\theta||_{\mathbb{R}^D} \in (r, r+\varepsilon)\}, \tag{3.5}$$

for $r, \varepsilon > 0$ to be chosen. To connect this to the notation in the preceding sections, the sets $\Theta_{r,\varepsilon}$ will play the role of the initialization (or starting) region $\mathcal{S}$, while B_s (for suitable s) corresponds to the target region $\mathcal{T}$ where the posterior mass concentrates. The 'intermediate' region $\mathcal{W} = \Theta_{s,\eta}$ representing the 'free-energy barrier' is constructed in the proofs of the theorems to follow.

Our results hold for general Markov chains whose invariant measure equals the posterior measure (3.2), and which admit a bound on their 'typical' step sizes. As step sizes can be random, this assumption needs to be accommodated in the probabilistic framework describing the transition probabilities of the chain. Let $\mathcal{P}_N(\theta, A), N \in \mathbb{N}$, (for $\theta \in \mathbb{R}^D$ and Borel sets $A \subseteq \mathbb{R}^D$), denote a sequence of Markov kernels describing the Markov chain dynamics employed for the computation of the posterior distribution $\Pi(\cdot|Z^{(N)})$. Recall that a probability measure μ on $\mathbb{R}^D$ is called invariant for $\mathcal{P}_N$ if $\int_{\mathbb{R}^D} \mathcal{P}_N(\theta, A)\mathrm{d}\mu(\theta) = \mu(A)$ for all Borel sets A.

Assumption 3.1. *Let $\mathcal{P}_N(\cdot,\cdot)$ be a sequence of Markov kernels satisfying the following:*

(*i*) *$\mathcal{P}_N(\cdot,\cdot)$ has invariant distribution $\Pi(\cdot|Z^{(N)})$ from (3.2).*
(*ii*) *For some fixed $c_0 > 0$ and for sequences $Q = Q_N > 0$, $\eta = \eta_N > 0$, with P_0^N-probability approaching 1 as $N \to \infty$,*

$$\sup_{\theta \in B_Q} \mathcal{P}_N\left(\theta, \left\{\vartheta : ||\theta - \vartheta||_{\mathbb{R}^D} \geq \frac{\eta}{2}\right\}\right) \leq \mathrm{e}^{-c_0 N}, \quad N \geq 1.$$

This assumption states that typical steps of the Markov chain are, with high probability (both under the law of the Markov chain and the randomness of the invariant 'target' measure), concentrated in an area of size $\eta/2$ around the current state θ, uniformly in a ball of radius Q around $\theta_0 = 0$. For standard MCMC algorithms (such as pCN, MALA) whose proposal steps are based on the discretization of some continuous-time diffusion process, such conditions can be checked, as we will show in the next section.

Theorem 3.2. *Let $D/N \simeq \kappa > 0$, consider the posterior (3.2) arising from the model (3.1) and a $\mathcal{N}(0, I_D/D)$ prior of density π, and let $\theta_0 = 0$. Then there exists $\mathscr{G}$ and a fixed constant $s \in (0, 1/3)$ for which the following statements hold true.*

(*i*) *The expected likelihood $\ell(\theta)$ is unimodal with mode 0, locally log-concave near 0, radially symmetric, Lipschitz continuous and monotonically decreasing in $||\theta||_{\mathbb{R}^D}$ on $\mathbb{R}^D$.*
(*ii*) *For any fixed $r > 0$, with high probability the log-likelihood $\ell_N(\theta)$ and the posterior density $\pi(\cdot|Z^{(N)})$ are monotonically decreasing in $||\theta||_{\mathbb{R}^D}$ on the set $\{\theta : ||\theta||_{\mathbb{R}^D} \geq r\}$.*
(*iii*) *We have that $\Pi(B_s|Z^{(N)}) \xrightarrow{N\to\infty} 1$ in probability.*
(*iv*) *There exists $\varepsilon > 0$ such that for any (sequence of) Markov kernels $\mathcal{P}_N$ on $\mathbb{R}^D$ and associated chains $(\vartheta_k : k \geq 1)$ that satisfy assumption 3.1 for some $c_0 > 0$, $Q = 1 + \varepsilon$, sequence $\eta_N \in (0, s)$ and all $N \geq 1$ large enough, we can find an initialization point $\vartheta_0 \in \Theta_{2/3,\varepsilon}$ such that with high probability (under the law of $Z^{(N)}$ and the Markov chain), the hitting time τ_{B_s} for ϑ_k to reach B_s (with s as in (iii)) is lower bounded as*

$$\tau_{B_s} \geq \exp\left(\min\{c_0, 1\}N/2\right).$$

The interpretation is that despite the posterior being strictly increasing in the radial variable $||\theta||_{\mathbb{R}^D}$ (at least for $||\theta||_{\mathbb{R}^D} > r$, any $r > 0$—note that maximizers of the posterior density

may deviate from the 'ground truth' $\theta_0 = 0$ by some asymptotically vanishing error, cf. also proposition 4.1), MCMC algorithms started in $\Theta_{2/3,\varepsilon}$ will still take an exponential time before visiting the region B_s where the posterior mass concentrates. This is true for small enough step size independently of D, N. The result holds also for ϑ_0 drawn from an absolutely continuous distribution on $\Theta_{2/3,\varepsilon}$ as inspection of the proof shows. Finally, we note that at the expense of more cumbersome notation, the above high probability results (and similarly in theorem 3.3) could be made non-asymptotic, in the sense that for all $\delta > 0$ all statements hold with probability at least $1 - \delta$ for all $N \geq N_0(\delta)$ large enough, where the dependency of N_0 on δ can be made explicit.

For 'ellipsoidally supported' α-regular priors (b), the idea is similar but the geometry of the problem changes as the prior now 'prefers' low-dimensional subspaces of $\mathbb{R}^D$, forcing the posterior closer towards the ground truth $\theta_0 = 0$. We show that if the step size is small compared to a scaling N^{-b} for $b > 0$ determined by α, then the same hardness phenomenon persists. Note that 'small' is only 'polynomially small' in N and hence algorithmic hardness does not come from exponentially small step sizes.

Theorem 3.3. *Let $D/N \simeq \kappa > 0$, consider the posterior (3.2) arising from the model (3.1) and a $\mathcal{N}(0, \Sigma_\alpha)$ prior of density π for some $\alpha > d/2$, and let $\theta_0 = 0$. Define $b = (\alpha/d) - (1/2) > 0$. Then there exists $\mathscr{G}$ and some fixed constant $s_b \in (0, 1/2)$ for which the following statements hold true.*

- (*i*) *The expected likelihood $\ell(\theta)$ is unimodal with mode 0, locally log-concave near 0, radially symmetric, Lipschitz continuous and monotonically decreasing in $||\theta||_{\mathbb{R}^D}$ on $\mathbb{R}^D$.*
- (*ii*) *For any fixed $r > 0$, with high probability $\ell_N(\theta)$ is radially symmetric and decreasing in $||\theta||_{\mathbb{R}^D}$ on the set $\{\theta : ||\theta||_{\mathbb{R}^D} \geq rN^{-b}\}$.*
- (*iii*) *Defining $s = s_b N^{-b}$, we have $\Pi(B_s|Z^{(N)}) \xrightarrow{N\to\infty} 1$ in probability.*
- (*iv*) *There exist positive constants $\varepsilon, C > 0$ and $\nu = \nu(\kappa, \alpha, d) > 0$ such that for any (sequence of) Markov kernels $\mathcal{P}_N$ on $\mathbb{R}^D$ and associated chains $(\vartheta_k : k \geq 1)$ that satisfy assumption 3.1 for some $c_0 > 0$, $Q = Q_N = C\sqrt{N}$, sequence $\eta = \eta_N \in (0, s_b N^{-b})$ and all $N \geq 1$ large enough, we can find an initialization point $\vartheta_0 \in \Theta_{N^{-b},\varepsilon N^{-b}}$ such that with high probability (under the law of $Z^{(N)}$ and the Markov chain), the hitting time τ_{B_s} for ϑ_k to reach B_s is lower bounded as*

$$\tau_{B_s} \geq \exp\big(\min\{c_0, \nu\}N/2\big).$$

Again, **(iv)** holds as well for ϑ_0 drawn from an absolutely continuous distribution on $\Theta_{N^{-b},\varepsilon N^{-b}}$. We also note that ε depends only on α, κ, d and the choice of $\mathscr{G}$ but not on any other parameters.

Remark 3.4. As opposed to theorem 3.2, due to the anisotropy of the prior density π, the posterior distribution is no longer radially symmetric in the preceding theorem, whence part **(ii)** differs from theorem 3.2. But a slightly weaker form of monotonicity of the posterior density $\pi(\cdot|Z^{(N)})$ still holds: the same arguments employed to prove part **(ii)** of theorem 3.2 show that $\pi(\cdot|Z^{(N)})$ is decreasing on $\{\theta : ||\theta||_{\mathbb{R}^D} \geq rN^{-b}\}$ (any $r > 0$) along the half-lines through 0, i.e.

$$\mathbb{P}_0^N\big(\pi(ve|Z^{(N)}) \leq \pi(v'e|Z^{(N)}) \quad \text{for all } v \geq v' \geq r,\ e \in \mathbb{R}^D, ||e||_{\mathbb{R}^D} = N^{-b}\big) \xrightarrow{N\to\infty} 1. \tag{3.6}$$

We note that this notion precludes the possibility of $\pi(\cdot|Z^{(N)})$ having extremal points outside of the region of dominant posterior mass, and implies that moving toward the origin will *always* increase the posterior density. As a result, many typical Metropolis–Hastings would be encouraged to accept such 'radially inward' moves, if they arise as a proposal. Thus, crucially, our exponential hitting time lower bound in part **(iv)** arises not through multi-modality, but merely through volumetric properties of high-dimensional Gaussian measures.

Remark 3.5 (On the step size condition). One may wonder whether larger step sizes can help to overcome the negative result presented in the last theorem. If the step sizes are 'time-homogeneous' and $\gg N^{-b}$ on average, then we may hit the region where the posterior is supported at some time. This would happen 'by chance' and not because the data (via ℓ_N) would

suggest to move there, and future proposals will likely be outside of that bulk region, so that the chain will either exit the relevant region again or become deterministic because an accept/reject step refuses to move into such directions. In this sense, a negative result for (polynomially) small step sizes gives fundamental limitations on the ability of the chain to explore the precise characteristics of the posterior distribution. We also remark that the Lipschitz constants of $\nabla\ell(\theta)$ are of order D or D^{1+b} in the preceding theorems, respectively. A Markov chain obtained from discretizing a continuous diffusion process (such as MALA discussed in the next section) will generally require step sizes that are inversely proportional to that Lipschitz constant in order to inherit the dynamics from the continuous process. For such examples, assumption 3.1 is natural. But as discussed at the end of the introduction, there exists a variety of 'non-local' MCMC algorithms for which this step size assumption may not be satisfied.

(b) Implications for common MCMC methods with 'cold-start'

The preceding general hitting time bounds apply to commonly used MCMC methods in high-dimensional statistics. We focus in particular on algorithms that are popular with PDE models and inverse problems, see, e.g. [50,51] and also [14] for many more references. We illustrate this for two natural examples with Metropolis–Hastings adjusted random walk and gradient algorithms. Other examples can be generated without difficulty.

(i) Preconditioned Crank–Nicolson

We first give some hardness results for the popular pCN algorithm. A dimension-free convergence analysis for pCN was given in the important paper by Hairer *et al.* [52] based on ideas from Hairer *et al.* [53]. The results in the present section show that while the mixing bounds from Hairer *et al.* [52] are in principle uniform in D, the implicit dependence of the constants on the conditions on the log-likelihood-function in [52] can re-introduce exponential scaling when one wants to apply the results from Hairer *et al.* [52] to concrete (N-dependent) posterior distributions. This confirms a conjecture about pCN made in Section 1.2.1 of Nickl & Wang [12].

Let $\mathcal{C}$ denote the covariance of some Gaussian prior on $\mathbb{R}^D$ with density π. Then the pCN algorithm for sampling from some posterior density $\pi(\theta|Z^{(N)}) \propto e^{\ell_N(\theta)}\pi(\theta)$ is given as follows. Let $(\xi_k : k \geq 1)$ be an i.i.d. sequence of $\mathcal{N}(0,\mathcal{C})$ random vectors. For initializer $\vartheta_0 \in \mathbb{R}^D$, step size $\beta > 0$ and $k \geq 1$, the MCMC chain is then given by

1. PROPOSAL: $p_k \sim \sqrt{1-\beta}\vartheta_{k-1} + \sqrt{\beta}\xi_k$,
2. ACCEPT–REJECT: Set

$$\vartheta_k = \begin{cases} p_k & \text{w.p. } \min\left\{1, e^{\ell_N(p_k)-\ell_N(\vartheta_{k-1})}\right\}, \\ \vartheta_{k-1} & \text{else.} \end{cases} \tag{3.7}$$

By standard Markov chain arguments one verifies (see [52] or Ch.1 in [14]) that the (unique) invariant density of $(\vartheta_k : k \geq 1)$ equals $\pi(\cdot|Z^{(N)})$.

We now give a hitting time lower bound for the pCN algorithm which holds true in the regression setting for which the main theorems 3.2 and 3.3 (for generic Markov chains) were derived. In particular, we emphasize that the lower bounds to follow hold for the choice of regression 'forward' map $\mathcal{G}$ constructed in the proofs of theorems 3.2 and 3.3. As for the general results, we treat the two cases of $\mathcal{C} = I_D/D$ or $\mathcal{C} = \Sigma_\alpha$ separately.

Theorem 3.6. *Let ϑ_k denote the pCN Markov chain from (3.7).*

(*i*) *Assume the setting of theorem 3.2 with $\mathcal{C} = I_D/D$, and let $\mathcal{G}$ be as in theorem 3.2. Then there exist constants $c_1, c_2, \varepsilon > 0$ such that for any $\beta \leq c_1$, there is an initialization point $\vartheta_0 \in \Theta_{2/3,\varepsilon}$ such that the hitting time $\tau_{B_s} = \inf\{k : \vartheta_k \in B_s\}$ (for B_s as in (3.4)) satisfies with high probability (under the law of the data and of the Markov chain) as $N \to \infty$ that $\tau_{B_s} \geq \exp(c_2 D)$.*

(*ii*) *Assume the setting of theorem 3.3 with* $\mathcal{C} = \Sigma_\alpha$ *for* $\alpha > d/2$, *and let* $\mathcal{G}$ *be as in theorem 3.2. Then there exist constants* $c_1, c_2, \varepsilon > 0$ *such that if* $\beta \le c_1 N^{-1-2b}$ *there is an initialization point* $\vartheta_0 \in \Theta_{N^{-b}, \varepsilon N^{-b}}$ *such that the hitting time* $\tau_{B_s} = \inf\{k : \vartheta_k \in B_s\}$ *satisfies with high probability that* $\tau_{B_s} \ge \exp(c_2 D)$.

(ii) Gradient-based Langevin algorithms

We now turn to *gradient-based* Langevin algorithms which are based on the discretization of continuous-time diffusion processes [10,50]. A polynomial time convergence analysis for the *unadjusted* Langevin algorithm in the strongly log-concave case has been given in [10,11] and also in [54] for the Metropolis-adjusted case (MALA). We show here that for unimodal but not globally log-concave distributions, the MCMC scheme can take an exponential time to reach the bulk of the posterior distribution. For simplicity we focus on the Metropolis-adjusted Langevin algorithm which is defined as follows. Let $(\xi_k : k \ge 1)$ be a sequence of i.i.d. $\mathcal{N}(0, I_D)$ variables, and let $\gamma > 0$ be a step size.

1. PROPOSAL: $p_k = \vartheta_{k-1} + \gamma \nabla \log \pi(\vartheta_{k-1} | Z^{(N)}) + \sqrt{2\gamma}\xi_k$.
2. ACCEPT–REJECT: Set

$$\vartheta_k = \begin{cases} p_k & \text{w.p. } \min\left\{1, \dfrac{\pi(\mathrm{p_k}|Z^{(N)}) \exp\left(-||\vartheta_{k-1} - \mathrm{p_k} - \gamma \nabla \log \pi(\mathrm{p_k}|Z^{(N)})||^2\right)}{\pi(\vartheta_{k-1}|Z^{(N)}) \exp\left(-||\mathrm{p_k} - \vartheta_{k-1} - \gamma \nabla \log \pi(\vartheta_{k-1}|Z^{(N)})||^2\right)}\right\} \\ \vartheta_{k-1} & \text{else.} \end{cases}, \tag{3.8}$$

Again, standard Markov chain arguments show that $\Pi(\cdot|Z^{(N)})$ is indeed the (unique) invariant distribution of $(\vartheta_k : k \ge 1)$. We note here that for the forward $\mathcal{G}$ featuring in our results to follows, $\nabla \log \pi$ may only be well-defined (Lebesgue-) almost everywhere on $\mathbb{R}^D$ due to our piece-wise smooth choice of w, see (4.6) below. However, since all proposal densities involved possess a Lebesgue density, this specification almost everywhere suffices in order to propagate the Markov chain with probability 1. Alternatively one could also straightforwardly avoid this technicality by smoothing our choice of function w in (4.6), which we refrain from for notational ease.

Theorem 3.7. *Let* ϑ_k *denote the MALA Markov chain from* (3.8).

(*i*) *Assume the setting of theorem 3.2, with* $\mathcal{N}(0, I_D/D)$ *prior, and let* $\mathcal{G}$ *also be as in theorem 3.2. There exists some* $c_1, c_2, \varepsilon > 0$ *such that if the step size of* $(\vartheta_k : k \ge 1)$ *satisfies* $\gamma \le c_1/N$, *then there is an initialization point* $\vartheta_0 \in \Theta_{2/3,\varepsilon}$ *such that the hitting time* $\tau_{B_s} = \inf\{k : \vartheta_k \in B_s\}$ *(for* B_s *as in* (3.4)*) satisfies with high probability (under the law of the data and of the Markov chain) as* $N \to \infty$ *that* $\tau_{B_s} \ge \exp(c_2 D)$.

(*ii*) *Assume the setting of theorem 3.3, with a* $\mathcal{N}(0, \Sigma_\alpha)$ *prior, and let* $\mathcal{G}$ *also be as in theorem 3.3. Then there exist some constant* $c_1, c_2, \varepsilon > 0$ *such that whenever* $\gamma \le c_1 N^{-1-b-2\alpha}$, *there is an initialization point* $\vartheta_0 \in \Theta_{N^{-b}, \varepsilon N^{-b}}$, *such that the hitting time* $\tau_{B_s} = \inf\{k : \vartheta_k \in B_s\}$ *satisfies with high probability (under the law of the data and of the Markov chain) that* $\tau_{B_s} \ge \exp(c_2 D)$.

As mentioned in remark 3.5, a bound on the step size that is inversely proportional to the Lipschitz constant of $\nabla \ell$ is natural for algorithms like MALA that arise from discretization of a continuous-time Markov process, see e.g. [11,54]. We emphasize again that these Lipschitz constants are D- and N-dependent, so that the required bounds on γ are not unnatural. 'Optimal' step size prescriptions for MALA [54–57] derived for Gaussian and log-concave targets or, more generally, mean-field limits (in which the posterior distribution possesses a product or mean-field structure, unlike in the models considered here) would need to be adjusted to our model classes to be comparable.

4. Proofs of the main theorems

We begin in §4*a* by constructing the family of regression maps $\mathcal{G}$ underlying our results from §3. Section 4*b*,*c* reduce the hitting time bounds from theorems 3.2 and 3.3 (for general Markov chains) to hitting time bounds for intermediate 'free energy barriers' that the Markov chain needs to travel through. Subsequently, theorems 3.3 and 3.2 are proved in §4*d*,*e*, respectively. Finally, the proofs for pCN (theorem 3.6) and MALA (theorem 3.7) are contained in §4*f*.

(a) Radially symmetric choices of G

We start with our parameterization of the map $\mathcal{G}$. In our regression model and since $\mathbb{E}\varepsilon^2 = 1$,

$$\ell(\theta) = -\frac{N}{2}\mathrm{E}^1_{\theta_0}|Y - \mathcal{G}(\theta)(X)|^2 = -\frac{N}{2}||\mathcal{G}(\theta_0) - \mathcal{G}(\theta)||^2_{L^2} - \frac{N}{2}, \quad \theta \in \mathbb{R}^D. \tag{4.1}$$

We have $\theta_0 = 0$ and by subtracting a fixed function $\mathcal{G}(0)$ from $\mathcal{G}(\theta)$ if necessary we can also assume that $\mathcal{G}(\theta_0) = 0$. In this case, since $\mathrm{vol}(\mathcal{X}) = 1$,

$$\ell(\theta) = -\frac{N}{2}||\mathcal{G}(\theta)||^2_{L^2} - \frac{N}{2}. \tag{4.2}$$

Take a bounded continuous function $w : [0, \infty] \to [0, ||w||_\infty)$ with a unique minimizer $w(0) = 0$ and take $\mathcal{G}$ of the 'radial' form

$$\mathcal{G}(\theta) = \sqrt{w(||\theta||_{\mathbb{R}^D})} \times g(x),\ \theta \in \mathbb{R}^D, x \in \mathcal{X},$$

where

$$g : \mathcal{X} \to [g_{\min}, g_{\max}],\ 0 < g_{\min} < g_{\max} < \infty,\ ||g||_{L^2_\mu(\mathcal{X})} = 1.$$

The assumption $\mathcal{G}(\theta_0) = 0$ implies $Y_i = 0 + \varepsilon_i$ under $P^N_{\theta_0}$, so that we have

$$\begin{aligned}\ell_N(\theta) &= -\frac{1}{2}\sum_{i=1}^N |\varepsilon_i - \sqrt{w(||\theta||)}g(X_i)|^2 \\ &= -\frac{w(||\theta||_{\mathbb{R}^D})}{2}\sum_{i=1}^N g^2(X_i) - \frac{1}{2}\sum_{i=1}^N \varepsilon_i^2 + \sqrt{w(||\theta||)}\sum_{i=1}^N \varepsilon_i g(X_i),\end{aligned} \tag{4.3}$$

and the average log-likelihood is

$$\ell(\theta) = \mathrm{E}^N_{\theta_0}\ell_N(\theta) = -\frac{N}{2}w(||\theta||_{\mathbb{R}^D}) - \frac{N}{2}, \theta \in \mathbb{R}^D. \tag{4.4}$$

Define ϵ-annuli of Euclidean space

$$\Theta_{r,\epsilon} = \left\{\theta \in \mathbb{R}^D : ||\theta||_{\mathbb{R}^D} \in (r, r+\epsilon)\right\},\ r \geq 0. \tag{4.5}$$

We then also set, for any $s \geq 0,\ \epsilon > 0$,

$$w_-(r,\epsilon) = \inf_{s\in(r,r+\epsilon)} w(s), \quad w_+(r,\epsilon) = \sup_{s\in(r,r+\epsilon)} w(s).$$

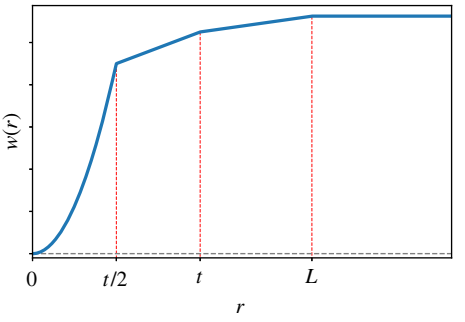
For our main theorems the map w will be monotone increasing and the preceding notation w_-, w_+ is then not necessary, but proposition 4.2 is potentially also useful in non-monotone settings (as remarked after its proof), hence the slightly more general notation here.

The choice that $\mathcal{G}$ is radial is convenient in the proofs, but means that the model is only identifiable up to a rotation for $\theta \neq 0$. One could easily make it identifiable by more intricate choices of $\mathcal{G}$, but the main point for our negative results is that the function ℓ has a unique mode at the ground truth parameter θ_0 and is identifiable there.

(i) A locally log-concave, globally monotone choice of w

Define for $t < L$ and any $r > 0$ the function $w : [0, \infty) \to \mathbb{R}$ as

$$\begin{aligned} w(r) = {} & 4(Tr)^2 1_{[0,t/2)} \\ & + [(Tt)^2 + T(r - t/2)]1_{[t/2,t)}(r) \\ & + [(Tt)^2 + (Tt/2) + \rho(r - t)]1_{[t,L)}(r) \\ & + [(Tt)^2 + (Tt/2) + \rho(L - t)]1_{[L,\infty)}(r), \end{aligned} \tag{4.6}$$



where $T > \rho$, are fixed constants to be chosen. Note that w is monotone increasing and

$$||w||_\infty = (Tt)^2 + \left(\frac{Tt}{2}\right) + \rho(L - t) < \infty. \tag{4.7}$$

The function w is quadratic near its minimum at the origin up until $t/2$, from when onwards it is piece-wise linear. In the linear regime it initially has a 'steep' ascent of gradient T until t, then grows more slowly with small gradient ρ from t until L, and from then on is constant. The function w is not C^∞ at the points $r = t/2$, $r = t$, $r = L$, but we can easily make it smooth by convolving with a smooth function supported in small neighbourhoods of its breakpoints r without changing the findings that follow. We abstain from this to simplify notation.

The following proposition summarizes some monotonicity properties of the empirical log-likelihood function arising from the above choice of w.

Proposition 4.1. *Let w be as in (4.6). Then there exists $C > 0$ such that for any $r_0 > 0$ and $N \geq 1$, we have*

$$P_0^N \left(\sup_{r_0 \leq r < s \leq L} \sup_{||\theta_s|| = s, ||\theta_r|| = r} \frac{\ell_N(\theta_s) - \ell_N(\theta_r)}{w(s) - w(r)} \leq -\frac{N}{4} \right) \geq 1 - \frac{C}{N} - \frac{C}{Nw(r_0)}.$$

In particular, if $r_0 < t/2$ is such that $(Tr_0)^2 N \to \infty$ as $N \to \infty$ then the r.h.s. is $1 - o(1)$.

Proof. Recalling (4.4), (4.3) and since w is monotonically increasing, we bound

$$\begin{aligned} & P_0^N \left(\ell_N(\theta_s) - \ell_N(\theta_r) > \left(\frac{N}{4}\right)(w(r) - w(s)) \right) \\ & = P_0^N \left(\ell_N(\theta_s) - \ell_N(\theta_r) - (\ell(\theta_s) - \ell(\theta_r)) > -\frac{N}{4}(w(r) - w(s)) \right) \\ & = \Pr \left(\frac{(w(r) - w(s))}{2} \sum_{i=1}^N (g^2(X_i) - 1) + \left(\sqrt{w(s)} - \sqrt{w(r)}\right) \sum_{i=1}^N \varepsilon_i g(X_i) > \frac{N}{4}(w(s) - w(r)) \right) \\ & = \Pr \left(-\sum_{i=1}^N (g^2(X_i) - \mathbb{E}g^2(X))/2 + \frac{1}{\sqrt{w(s)} + \sqrt{w(r)}} \sum_{i=1}^N \varepsilon_i g(X_i) > \frac{N}{4} \right) \\ & \leq \Pr \left(\left| \sum_{i=1}^N (g^2(X_i) - \mathbb{E}g^2(X)) \right| > \frac{N}{4} \right) + \Pr \left(|\sum_{i=1}^N \varepsilon_i g(X_i)| > \frac{2N\sqrt{w(r_0)}}{8} \right) \\ & = \mathcal{O}\left(\frac{1}{N}\right) + \mathcal{O}\left(\frac{1}{Nw(r_0)}\right), \end{aligned}$$

using Chebyshev's inequality in the last step. Since the events in the penultimate step do not depend on $r < s \in [r_0, L]$, the result follows. ∎

royalsocietypublishing.org/journal/rsta Phil. Trans. R. Soc. A 381: 20220150

(b) Bounds for posterior ratios of annuli

A key quantity in the proofs to follow will be to obtain asymptotic ($N \to \infty$) bounds of the following functional (recalling the definition of the Euclidean annuli $\Theta_{r,\varepsilon}$ from (4.5)),

$$\mathscr{F}_N(r,\varepsilon) = \frac{1}{N}\log \int_{\Theta_{r,\varepsilon}} \mathrm{e}^{\ell_N(\theta)}\, \mathrm{d}\Pi(\theta), \quad r \geq 0,\ \varepsilon > 0, \tag{4.8}$$

in terms of the map w. As a side note, we remark that this functional has a long history in the statistical physics of glasses, in which it is often referred to as the *Franz–Parisi* potential [7,58].

Proposition 4.2. *Consider the regression model (3.1) with radially symmetric choice of $\mathscr{G}$ from §4a such that $||w||_\infty \leq W$ for some fixed $W < \infty$ (independent of D, N), and let $\Pi = \Pi_N$ denote a sequence of prior probability measures on $\mathbb{R}^D$.*

(*i*) *Suppose that for some radii $0 < s < \sigma$, constants $\varepsilon, \eta, \nu > 0$ and for all $N \geq 1$ large enough, we have*

$$\frac{1}{N}\log \frac{\Pi(\Theta_{s,\eta})}{\Pi(\Theta_{\sigma,\varepsilon})} \leq -2\nu - \frac{(w_+(\sigma,\varepsilon) - w_-(s,\eta))}{2}. \tag{4.9}$$

Then the posterior distribution $\Pi(\cdot|Z^{(N)})$ from (3.2) arising in the model (3.1) satisfies that with high P_0^N-probability as $N \to \infty$,

$$\frac{\Pi(\Theta_{s,\eta}|Z^{(N)})}{\Pi(\Theta_{\sigma,\varepsilon}|Z^{(N)})} \leq \mathrm{e}^{-\nu N}. \tag{4.10}$$

(*ii*) *If in addition w is monotone increasing on $[0,\infty)$ and if for some $Q > 1 + \varepsilon$,*

$$\frac{1}{N}\log \frac{\Pi(B_Q^c)}{\Pi(\Theta_{\sigma,\varepsilon})} \leq -2\nu, \tag{4.11}$$

then the posterior distribution $\Pi(\cdot|Z^{(N)})$ also satisfies (with high probability as $N \to \infty$) that

$$\frac{\Pi(B_Q^c|Z^{(N)})}{\Pi(\Theta_{\sigma,\varepsilon}|Z^{(N)})} \leq \mathrm{e}^{-\nu N}. \tag{4.12}$$

Remark 4.3 (The prior condition for w from (4.6)). If $\sigma > s > t$, for w from (4.6), the 'likelihood' term in proposition 4.2 is

$$\frac{w_+(\sigma,\varepsilon)}{2} - \frac{w_-(s,\eta)}{2} \leq \frac{\rho(\sigma+\varepsilon-t) - \rho(s-t)}{2} = \frac{\rho}{2}(\sigma+\varepsilon-s) > 0, \tag{4.13}$$

so that if we also assume

$$Tt + \rho L = \mathcal{O}(\sqrt{N}), \tag{4.14}$$

to control ω_N, ω_N' in the proof that follows, then to verify (4.9) it suffices to check

$$\frac{1}{N}\log \frac{\Pi(\Theta_{s,\eta})}{\Pi(\Theta_{\sigma,\varepsilon})} \leq -2\nu - \frac{\rho}{2}(\sigma+\varepsilon-s), \tag{4.15}$$

for all large enough N.

Proof. Proof of part (i). From the definition of ℓ_N in (4.3) we first note that for all $r \geq 0, \epsilon > 0$,

$$\inf_{\theta \in \Theta_{r,\epsilon}} \ell_N(\theta) \geq -\frac{1}{2}\sum_{i=1}^N \varepsilon_i^2 - \frac{1}{2} w_+(r,\epsilon) \sum_{i=1}^N g^2(X_i) - \sqrt{w_+(r,\epsilon)} \left| \sum_{i=1}^N \varepsilon_i g(X_i) \right|$$

and

$$\sup_{\theta \in \Theta_{r,\epsilon}} \ell_N(\theta) \leq -\frac{1}{2}\sum_{i=1}^N \varepsilon_i^2 - \frac{1}{2} w_-(r,\epsilon) \sum_{i=1}^N g^2(X_i) + \sqrt{w_+(r,\epsilon)} \left| \sum_{i=1}^N \varepsilon_i g(X_i) \right|.$$

We can now further bound, for our $\mathcal{G}$,

$$\frac{1}{N} \log \int_{\Theta_{r,\epsilon}} \mathrm{e}^{\ell_N(\theta)}\, \mathrm{d}\Pi(\theta) \leq -\frac{1}{2N}\sum_{i=1}^N \varepsilon_i^2 - \frac{w_-(r,\epsilon)}{2N} \sum_{i=1}^N g^2(X_i) + \frac{\sqrt{w_+(r,\epsilon)}}{N} \left| \sum_{i=1}^N \varepsilon_i g(X_i) \right| + \frac{\log \Pi(\Theta_{r,\epsilon})}{N}$$

and

$$\frac{1}{N} \log \int_{\Theta_{r,\epsilon}} \mathrm{e}^{\ell_N(\theta)} \mathrm{d}\Pi(\theta) \geq -\frac{1}{2N}\sum_{i=1}^N \varepsilon_i^2 - \frac{w_+(r,\epsilon)}{2N} \sum_{i=1}^N g^2(X_i) - \frac{\sqrt{w_+(r,\epsilon)}}{N} \left| \sum_{i=1}^N \varepsilon_i g(X_i) \right| + \frac{\log \Pi(\Theta_{r,\epsilon})}{N}.$$

We estimate $\sqrt{w_+(r,\epsilon)} \leq \bar{w}(r,\epsilon) = \max(w_+(r,\epsilon), 1)$, and noting that

$$\mathbb{E}\varepsilon_i^2 = 1 = \mathbb{E}g^2(X_i) \quad \text{and} \quad \mathbb{E}\varepsilon_i g(X_i) = 0,$$

we can use Chebyshev's (or Bernstein's) inequality to construct an event of high probability such that the functional $\mathscr{F}_N$ from (4.8) is bounded as

$$\mathscr{F}_N(r,\epsilon) \leq -\frac{1}{2} - \frac{w_-(r,\epsilon)}{2} + \frac{\log \Pi(\Theta_{r,\epsilon})}{N} + \omega_N(s,\eta) \tag{4.16}$$

and

$$\mathscr{F}_N(r,\epsilon) \geq -\frac{1}{2} - \frac{w_+(r,\epsilon)}{2} + \frac{\log \Pi(\Theta_{r,\epsilon})}{N} + \omega'_N(r,\epsilon), \tag{4.17}$$

where

$$\omega_N(r,\epsilon) = \mathcal{O}\Big(1 + \frac{w_-(r,\epsilon) + \bar{w}(r,\epsilon)}{\sqrt{N}}\Big),\ \omega'_N(s) = \mathcal{O}\Big(1 + \frac{w_+(r,\epsilon) + \bar{w}(r,\epsilon)}{\sqrt{N}}\Big), \tag{4.18}$$

and this is uniform in all (r,ϵ) since $||w||_\infty \leq W$ is bounded. Using the above with (r,ϵ) chosen as (s,η) and (σ,ε) respectively, we then obtain

$$\begin{aligned}
\frac{1}{N} \log \frac{\Pi(\Theta_{s,\eta}|Z^{(N)})}{\Pi(\Theta_{\sigma,\varepsilon}|Z^{(N)})} &= \mathscr{F}_N(s,\eta) - \mathscr{F}_N(\sigma,\varepsilon) \\
&\leq -\frac{w_-(s,\eta)}{2} + \frac{\log \Pi(\Theta_{s,\eta})}{N} + \frac{w_+(\sigma,\varepsilon)}{2} - \frac{\log \Pi(\Theta_{\sigma,\varepsilon})}{N} + \omega_N(s,\eta) - \omega'_N(\sigma,\varepsilon) \\
&= -\frac{w_-(s,\eta)}{2} + \frac{w_+(\sigma,\varepsilon)}{2} + \frac{1}{N} \log \frac{\Pi(\Theta_{s,\eta})}{\Pi(\Theta_{\sigma,\varepsilon})} + \omega_N(s,\eta) - \omega'_N(\sigma,\varepsilon),
\end{aligned} \tag{4.19}$$

with high $P_{\theta_0}^N$-probability. The result now follows from the hypothesis (4.9) and since the terms ω_N, ω'_N are $\mathcal{O}(1)$.

(Proof of part ii). The proof of part **(ii)** follows from an obvious modification of the previous arguments. ∎

In the case where $\Pi(\Theta_{s,\eta})$ and $\Pi(\Theta_{\sigma,\varepsilon})$ are comparable (so that the l.h.s. in (4.9) converges to zero), a local optimum at σ in the function w away from zero can verify the last inequality for 'intermediate' s such that $w(s) - w(\sigma) \leq -2\nu$. This can be used to give computational hardness

results for MCMC of multi-modal distributions. But we are interested in the more challenging case of 'unimodal' examples w from (4.6). Before we turn to this, let us point out what can be said about the hitting times of Markov chains if the conclusion (4.10) of proposition 4.2 holds.

(c) Bounds for Markov chain hitting times

(i) Hitting time bounds for intermediate sets $\Theta_{s,\eta}$

In (4.10), we can think of $\Theta_{\sigma,\varepsilon}$ as the 'initialization region' (further away from θ_0) and $\Theta_{s,\eta}$ for intermediate s is the 'barrier' before we get close to $\theta_0 = 0$. The last bound permits the following classic hitting time argument, taken from Ben Arous *et al.* [5], see also [8].

Proposition 4.4. *Consider any Markov chain $(\vartheta_k : k \in \mathbb{N})$ with invariant measure $\mu = \Pi(\cdot|Z^{(N)})$ for which (4.10) holds. For constants $\eta < \sigma - s$, suppose ϑ_0 is started in $\Theta_{\sigma,\varepsilon}$, $\mu(\Theta_{\sigma,\varepsilon}) > 0$, drawn from the conditional distribution $\mu(\cdot|\Theta_{\sigma,\varepsilon})$, and denote by τ_s the hitting time of the Markov chain onto $\Theta_{s,\eta}$, that is, the number τ_s of iterates required until ϑ_k visits the set $\Theta_{s,\eta}$. Then*

$$\Pr(\tau_s \le K) \le K\,\mathrm{e}^{-\nu N}, \quad K > 0.$$

Similarly, on the event where (4.12) holds we have that

$$\Pr(\tau_{B_L^c} \le K) \le K\,\mathrm{e}^{-\nu N}, \quad K > 0.$$

Proof of proposition 4.4. We have

$$\begin{aligned}\Pr(\tau_s \le K) &= \Pr(\vartheta_k \in \Theta_{s,\eta} \quad \text{for some } 1 \le k \le K | \vartheta_0 \in \Theta_{\sigma,\varepsilon}) \\ &= \frac{\Pr(\vartheta_0 \in \Theta_{\sigma,\varepsilon}, \vartheta_k \in \Theta_{s,\eta} \quad \text{for some } 1 \le k \le K)}{\mu(\Theta_{\sigma,\varepsilon})} \le \frac{\sum_{k\le K} \Pr(\vartheta_k \in \Theta_{s,\eta})}{\mu(\Theta_{\sigma,\varepsilon})} \\ &\le K \frac{\mu(\Theta_{s,\eta})}{\mu(\Theta_{\sigma,\varepsilon})} \le K\,\mathrm{e}^{-\nu N}.\end{aligned}$$

The second claim is proved analogously. ∎

The last proposition holds 'on average' for initializers $\vartheta_0 \sim \mu(\cdot|\Theta_{\sigma,\varepsilon})$, and since $\Pr = \mathbb{E}_{\mu(\cdot|\Theta_{\sigma,\varepsilon})} \Pr_{\vartheta_0}$ where $\Pr_{\vartheta_0}$ is the law of the Markov chain started at ϑ_0, the hitting time inequality holds at least for one point in $\Theta_{\sigma,\varepsilon}$ since $\inf_{\vartheta_0} \Pr_{\vartheta_0} \le \mathbb{E}_{\mu(\cdot|\Theta_{\sigma,\varepsilon})} \Pr_{\vartheta_0}$.

(ii) Reducing hitting times for B_s to ones for $\Theta_{s,\eta}$

We now reduce part **(iv)** of theorems 3.2 and 3.3, i.e. bounds on the hitting time of the region B_s in which the posterior contracts, to a bound for the hitting time τ_s for the annulus $\Theta_{s,\eta}$, which is controlled in proposition 4.4. To this end, in the case of theorem 3.2, we suppose that propositions 4.2 and 4.4 are verified with $\nu = 1\sigma = 2/3$ some $\varepsilon > 0$ and Q, s, η as in the theorem, and in the case of theorem 3.3, we assume the same with choice $\sigma = N^{-b}$ and $\nu > 0$ given after (4.27) below. For c_0 from assumption 3.1, define the events

$$A_N := \left\{ \forall k \le \mathrm{e}^{(\nu \wedge c_0)N/2} : ||\vartheta_{k+1} - \vartheta_k||_{\mathbb{R}^D} \le \frac{\eta}{2} \right\}.$$

We can then estimate, using assumption 3.1, that on the frequentist event on which proposition 4.4 holds (which we apply with $K = \mathrm{e}^{(\nu\wedge c_0)N/2} \le \mathrm{e}^{\nu N/2}$), under the probability law of the Markov chain we have

$$\begin{aligned}&\Pr(\tau_{B_s} \le \mathrm{e}^{(\nu\wedge c_0)N/2}) \le \Pr(\tau_{B_s} \le \mathrm{e}^{(\nu\wedge c_0)N/2}, A_N) + \Pr(A_N^c) \\ &\le \Pr(\tau_s \le \mathrm{e}^{(\nu\wedge c_0)N/2}) + \Pr(A_N^c, \tau_{B_Q^c} > \mathrm{e}^{(\nu\wedge c_0)N/2}) + \Pr(\tau_{B_Q^c} \le \mathrm{e}^{(\nu\wedge c_0)N/2}) \\ &\le 2\,\mathrm{e}^{-\nu N/2} + \mathrm{e}^{(\nu\wedge c_0)N/2} \sup_{\theta \in B_Q} \mathcal{P}_N\left(\theta, \left\{\vartheta : ||\theta - \vartheta||_{\mathbb{R}^D} \ge \frac{\eta}{2}\right\}\right) \\ &\le 2\,\mathrm{e}^{-(\nu\wedge c_0)N/2} + \mathrm{e}^{(\nu\wedge c_0)N/2 - c_0 N} \le 3\,\mathrm{e}^{-(\nu\wedge c_0)N/2},\end{aligned}$$

where in the second inequality we have used that on the events A_N, the Markov chain ϑ_k, when started in $\Theta_{2/3,\varepsilon}$, needs to pass through $\Theta_{s,\eta}$ in order to reach B_s.

(d) Proof of theorem 3.3

In this section, we use the results derived in the previous part of §4 to finish the proof of theorem 3.3. Parts **(i)** and **(ii)** of the theorem follow from proposition 4.1 and our choice of w in (4.6). We therefore concentrate on the proofs of part **(iii)** and **(iv)**. We start with proving a key lemma on small ball estimates for truncated α-regular Gaussian priors.

(i) Small ball estimates for α-regular priors

Let us first define precisely the notion of α-regular Gaussian priors. For some fixed $\alpha > d/2$, the prior Π arises as the truncated law $Law(\theta)$ of an α-regular Gaussian process with RKHS $\mathcal{H} = H^\alpha$, a Sobolev space over some bounded domain/manifold $\mathcal{X}$, see e.g. section 6.2.1 in [14] for details. Equivalently (under the Parseval isometry) we take a Gaussian Borel measure on the usual sequence space $\ell_2 \simeq L^2$ with RKHS equal to

$$h^\alpha = \left\{ (\theta_i)_{i=1}^\infty : \sum_{i=1}^\infty i^{2(\alpha/d)} \theta_i^2 = ||\theta||_{H^\alpha}^2 < \infty \right\}, \quad \alpha > \frac{d}{2}.$$

The prior Π is the truncated law of $\theta_D = (\theta_1, \ldots, \theta_D), D \in \mathbb{N}$.

Lemma 4.5. *Fix $z > 0$, $\alpha > d/2$ and $\kappa > 0$, and set*

$$b = \frac{\alpha}{d} - \frac{1}{2}, \quad \tau = \frac{1}{b} = \frac{2d}{2\alpha - d}.$$

Then if $D/N \simeq \kappa > 0$, there exist constants $\bar{c}_0 > c_0$ (depending on b, κ) such that for all N ($\geq N_0(z,b)$) large enough:

$$c_0(z + \kappa^{-\alpha/d} z^{-\tau/2})^{-\tau} \leq -\frac{1}{N} \log \Pi(||\theta||_{\mathbb{R}^D} \leq z N^{-b}) \leq \bar{c}_0 z^{-\tau}. \tag{4.20}$$

Proof of lemma 4.5. Note first that the L^2-covering numbers of the ball $h(\alpha, B)$ of radius B in H^α satisfy the well-known two-sided estimate

$$\log \mathcal{N}(\delta, ||\cdot||_{L^2}, h(\alpha, B)) \simeq \left(\frac{AB}{\delta} \right)^{d/\alpha}, \quad 0 < \delta < AB \tag{4.21}$$

for equivalence constants in $\simeq$ depending only on d, α. The upper bound is given in proposition 6.1.1 in [14] and a lower bound can be found as well in the literature [59] (by injecting $H^\alpha(\mathcal{X}_0)$ into $\tilde{H}^\alpha(\mathcal{X})$ for some strict sub-domain $\mathcal{X}_0 \subset \mathcal{X}$, and using metric entropy lower bounds for the injection $H^\alpha(\mathcal{X}_0) \to L^2(\mathcal{X}_0)$).

Using the results about small deviation asymptotics for Gaussian measures in Banach space [60]—specifically theorem 6.2.1 in [14] with $a = 2d/(2\alpha - d)$—and assuming $\alpha > d/2$, this means that the concentration function of the 'untruncated prior' satisfies the two-sided estimate

$$-\log \Pi(||\theta||_{L^2} \leq \gamma) \simeq \gamma^{-(2d/(2\alpha - d))} = \gamma^{-\tau}, \quad \gamma \to 0. \tag{4.22}$$

Here, restricting to $\gamma \in (0,1)$, the two-sided equivalence constants depend only on α, d. Setting

$$\gamma = z N^{-b}, \quad z > 0, \tag{4.23}$$

and noting that $b\tau = 1$, we hence obtain that for some constants $c_l, c_u > 0$,

$$e^{-c_l z^{-\tau} N} \leq \Pi(||\theta||_{L^2} \leq z N^{-b}) \leq e^{-c_u z^{-\tau} N}, \quad \text{any } z > 0. \tag{4.24}$$

We now show that as long as $D/N \approx \kappa > 0$, one may use the above asymptotics to derive the desired small ball probabilities for the projected prior on $\mathbb{R}^D$.

We obviously have, by set inclusion and projection,

$$\Pi(||\theta||_{\mathbb{R}^D} \le zN^{-b}) \ge \Pi(||\theta||_{L^2} \le zN^{-b}),$$

and hence it only remains to show the first inequality in equation (4.20). The Gaussian isoperimetric theorem (theorem 2.6.12 in [61]) and (4.24) imply that for $m \ge 4\sqrt{c_l}$ and some $c > 0$, we have that (with Φ denoting the c.d.f. for $\mathcal{N}(0,1)$)

$$\begin{aligned}
&\Pi\big(\theta = \theta_1 + \theta_2, ||\theta_1||_{L^2} \le zN^{-b}, ||\theta_2||_{h^\alpha} \le mz^{-\tau/2}\sqrt{N}\big)\\
&\ge \Phi\big(\Phi^{-1}\big(\Pi(\{\theta : ||\theta|| \le zN^{-b}\})\big) + mz^{-\tau/2}\sqrt{N}\big)\\
&\ge \Phi\big(-\sqrt{2c_l}z^{-\tau/2}\sqrt{N} + mz^{-\tau/2}\sqrt{N}\big) \ge 1 - \mathrm{e}^{-cz^{-\tau}N},
\end{aligned}$$

(see also the proof of lemma 5.17 in [19] for a similar calculation). Then if the event in the last probability is denoted by I we have

$$\Pi(||\theta_D||_{\mathbb{R}^D} \le zN^{-b}) \le \Pi(||\theta_D||_{\mathbb{R}^D} \le zN^{-b}, I) + \mathrm{e}^{-cz^{-\tau}N}.$$

On I, if $D/N \to \kappa > 0$ and by the usual tail estimate for vectors in h^α, we have for some $c' > 0$ the bound

$$||\theta - \theta_D||_{L^2} \le ||\theta_1||_{L^2} + c'D^{-\alpha/d}z^{-\tau/2}\sqrt{N} \le zN^{-b} + c'\kappa^{-\alpha/d}z^{-\tau/2}N^{-b},$$

so that for any $z > 0$,

$$\begin{aligned}
&\Pi(||\theta_D||_{\mathbb{R}^D} \le zN^{-b}) \le \Pi(||\theta||_{L^2} \le zN^{-b} + ||\theta - \theta_D||_{L^2}, I) + \mathrm{e}^{-cz^{-\tau}N}\\
&\le \Pi(||\theta||_{L^2} \le (2z + c'\kappa^{-\alpha/d}z^{-\tau/2})N^{-b}) + \mathrm{e}^{-cz^{-\tau}N}\\
&\le \mathrm{e}^{-c_u(2z+c'\kappa^{-\alpha/d}z^{-\tau/2})^{-\tau}N} + \mathrm{e}^{-cz^{-\tau}N},
\end{aligned}$$

and hence the lemma follows by appropriately choosing $c_0 > 0$. ∎

Remark 4.6. For statistical consistency proofs in nonlinear inverse problems, often *rescaled* Gaussian priors are used to provide additional regularization [12,13,19]. For these priors a computation analogous to the previous lemma is valid: specifically if we rescale θ by $\sqrt{N}\delta_N$, where $\delta_N = N^{-\alpha/(2\alpha+d)}$ so that $\sqrt{N}\delta_N = N^{(d/2)/(2\alpha+d)} = N^k$, then we just take $N^{-\beta+k} = N^{-b}$ in the above small ball computation, that is $-b = -\beta + k$ or $b = \beta - k$, and the same bounds (as well as the proof to follow) apply.

(ii) Proof of theorem 3.3, part (iv)

Lemma 4.5 and the hypotheses on η immediately imply

$$\begin{aligned}
&\Pi(\theta \in \Theta_{s,\eta}) = \Pi\big(||\theta||_{\mathbb{R}^D} \in (s_bN^{-b}, s_bN^{-b} + \eta)\big)\\
&\le \Pi\big(||\theta||_{\mathbb{R}^D} \le 2s_bN^{-b}\big) \le \mathrm{e}^{-c_0N(2s_b+\kappa^{-\alpha/d}(2s_b)^{-\tau/2})^{-\tau}}.
\end{aligned}$$

To lower bound $\Pi(\Theta_{N^{-b},\varepsilon N^{-b}})$, we choose ε large enough such that

$$\bar{c}_0(1+\varepsilon)^{-\tau} < c_0(1+\kappa^{-\alpha/d})^{-\tau},$$

which implies for all N large enough that

$$\begin{aligned}
\Pi\big(||\theta||_{\mathbb{R}^D} \in (N^{-b}, (1+\varepsilon)N^{-b})\big) &= \Pi(||\theta||_{\mathbb{R}^D} \le (1+\varepsilon)N^{-b}) - \Pi(||\theta||_{\mathbb{R}^D} \le N^{-b}))\\
&\ge \mathrm{e}^{-\bar{c}_0(1+\varepsilon)^{-\tau}N} - \mathrm{e}^{-c_0(1+\kappa^{-\alpha/d})^{-\tau}N}\\
&\ge \mathrm{e}^{-2\bar{c}_0(1+\varepsilon)^{-\tau}N}. \qquad (4.25)
\end{aligned}$$

Now, for w from (4.6), we set

$$t = t_bN^{-b}, \quad \rho \in (0,1],\ 0 < t_b < s_b < \frac{1}{2} < L < \infty,\ T = T_bN^b, \qquad (4.26)$$

for T_b to be chosen and ρ, L, s_b, t_b fixed constants, so that $||w||_\infty$ is bounded (uniformly in N) by a constant which depends only on T_b, L, ρ, whence (4.14) holds. Now the key inequality (4.15) with $s = s_b N^{-b}$ and with our choice of $\eta, \varepsilon, \sigma = N^{-b}$ will be satisfied if

$$c_0(2s_b + \kappa^{-\alpha/d}(2s_b)^{-\tau/2})^{-\tau} \geq 2\bar{c}_0(1+\varepsilon)^{-\tau} + 2\nu + \frac{\rho}{2}N^{-b}(1+\varepsilon - s_b). \tag{4.27}$$

We define ν to equal to 1/3 of the l.h.s. so that (4.27) will follow for the given s_b, κ, α, d by choosing ε large enough and whenever N is large enough.

Finally, let us note that with $Q = C\sqrt{N}$ for some $C \geq 2\mathbb{E}[||\theta||_{\ell_2}]$, where θ is the infinite Gaussian vector with RKHS h^α, we can deduce from theorem 2.1.20 and exercise 2.1.5 in [61] that

$$\Pr(||\theta||_{\mathbb{R}^D} \geq Q) \leq 2\exp(-c\,C^2 N/2), \quad \text{some } c > 0.$$

Thus, using also (4.25), choosing C large enough verifies (4.11). Since (4.25) and the a.s. boundedness of $\sup_\theta |\ell_N(\theta)|$ for ℓ_N from (4.3) imply that $\Pi(\Theta_{N^{-b},\varepsilon N^{-b}}|Z^{(N)}) > 0$ a.s., proposition 4.2 and then also proposition 4.4 apply for this prior, and the arguments from §4*c*(ii) yield the desired result.

(iii) Proof of theorem 3.3, part (iii)

We finish the proof of the theorem by showing point **(iii)** . We use the setting and choices from the previous section. Let us write $\mathbb{G}(A) = \int_A \mathrm{e}^{\ell_N(\theta)}\,\mathrm{d}\Pi(\theta)$ for any measurable set A. Recall the notation $B_r = \{\theta : ||\theta||_{\mathbb{R}^D} \leq r\}, r > 0$. Repeating the argument leading to (4.17) with $B_{t/2}$ in place of $\Theta_{r,\epsilon}$, and using lemma 4.5, we have with high probability

$$\frac{1}{N}\log \mathbb{G}(B_{t/2}) \geq -\frac{1}{2} - \frac{\sup_{r \leq t_b N^{-b}/2} w(r)}{2} - \bar{c}_0\left(\frac{t_b}{2}\right)^{-\tau} + \omega_N'(t/2),$$

where $\omega_N'(t/2) = \mathcal{O}(||w||_\infty/\sqrt{N}) = o(1)$. Likewise, we also have

$$\frac{1}{N}\log \mathbb{G}(B_s^c) \leq -\frac{1}{2} - \frac{\inf_{r \geq s_b N^{-b}} w(r)}{2} + \frac{1}{N}\log \Pi(B_s^c) + \omega_N''(s),$$

where $\omega_N''(s) = \mathcal{O}(||w||_\infty/\sqrt{N}) = o(1)$. We can assume that $\mathbb{G}(B_s^c) > 0$. Hence, since $\Pi(B_s^c) \to 1$ in view of lemma 4.5,

$$\begin{aligned}\frac{1}{N}\log\frac{\mathbb{G}(B_{t/2})}{\mathbb{G}(B_s^c)} &\geq -\frac{(Tt)^2}{2} - c_0\left(\frac{t_b}{2}\right)^{-\tau} + \frac{(Tt)^2 + (Tt/2) + \rho(s-t)}{2} + \frac{1}{N}\log \Pi(B_s^c) + o(1)\\ &\geq \frac{T_b t_b}{4} - c_0\left(\frac{t_b}{2}\right)^{-\tau} + o(1).\end{aligned} \tag{4.28}$$

Now, for $t_b < s_b$ fixed we can choose T_b large enough such that the last quantity exceeds 1 with high probability (in particular this retrospectively justifies the last $o(1)$ as then $||w||_\infty = \mathcal{O}(1)$ for our choice of T_b). Therefore, again with high probability

$$\frac{\mathbb{G}(B_{t/2})}{\mathbb{G}(B_s^c)} \geq \mathrm{e}^N \times (1 + o(1)). \tag{4.29}$$

For $M_{t,s} = \{\theta : t/2 < ||\theta||_{\mathbb{R}^D} \leq s\}$ this further implies that with high probability

$$\frac{\mathbb{G}(B_{t/2}) + \mathbb{G}(M_{t,s})}{\mathbb{G}(B_s^c)} \geq \mathrm{e}^N \times (1 + o(1)),$$

and then,

$$\begin{aligned}\Pi(B_s|Z^{(N)}) &= \frac{\mathbb{G}(B_{t/2}) + \mathbb{G}(M_{t,s})}{\mathbb{G}(B_{t/2}) + \mathbb{G}(M_{t,s}) + \mathbb{G}(B_s^c)}\\ &= \frac{\mathbb{G}(B_{t/2}) + \mathbb{G}(M_{t,s})}{(\mathbb{G}(B_{t/2}) + \mathbb{G}(M_{t,s}))(1 + (\mathbb{G}(B_s^c)/\mathbb{G}(B_{t/2})) + \mathbb{G}(M_{t,s}))} \to 1,\end{aligned}$$

again with high probability, which is what we wanted to show.

Remark 4.7. If the map w is globally convex, say $w(s) = Ts^2/2$ for all $s > 0$, then a 'large enough' choice of ε after (4.27) is not possible. It is here where global log-concavity of the likelihood function helps, as it enforces a certain 'uniform' spread of the posterior across its support via a global coercivity constant T. By contrast the above example of w is not convex, rather it is very spiked on $(0, t/2)$ and then 'flattens out'.

(e) Proof of theorem 3.2

The proof of theorem 3.2 proceeds along the same lines as that of theorem 3.3, with scaling t, L, ρ, s, η constant in N, corresponding to $b = 0$ in N^{-b}, and replacing the volumetric lemma 4.5 by the following basic result.

Lemma 4.8. *Let $\theta \sim \mathcal{N}(0, I_D/D)$. Let $a \in (0, 1/2)$. Then for all $D \geq D_0(a)$ large enough,*

$$-\frac{1}{D}\log \Pi(||\theta||_{\mathbb{R}^D} \leq z) \geq \frac{1}{2}\left(\frac{z^2}{2} - \log z - \frac{1}{2}\right), \quad \text{any } z \in (0, 1-a). \tag{4.30}$$

A proof of (4.30) is sketched in appendix B. As a consequence of the previous lemma

$$\frac{1}{N}\log \Pi(\Theta_{s,\eta}) \leq \frac{1}{N}\log \Pi(B_{2s}) \leq \frac{\kappa}{2}\left(\log 2s - 2s^2 + \frac{1}{2}\right).$$

Moreover, to lower bound $\Pi(\Theta_{2/3,\varepsilon})$, we choose $\varepsilon > 2/3$. Then, using theorem 2.5.7 in [61] as well as $\mathbb{E}||\theta|| \leq \mathbb{E}(||\theta||^2)^{1/2} = 1$, and then also (4.30) with $z = 2/3$, we obtain that

$$\begin{aligned}
\Pi(\Theta_{2/3,\varepsilon}) &\geq \Pi\left(|||\theta||_{\mathbb{R}^D} - 1| \leq \frac{1}{3}\right) \\
&\geq 1 - \Pi\left(||\theta||_{\mathbb{R}^D} \geq \mathbb{E}||\theta||_{\mathbb{R}^D} + \frac{1}{3}\right) - \Pi\left(||\theta||_{\mathbb{R}^D} \leq \frac{2}{3}\right) \\
&\geq 1 - \exp(-D/18) - \exp(-cD),
\end{aligned}$$

for some fixed constant $c > 0$ given by (4.30), whence $\Pi(\Theta_{2/3,\varepsilon}) \to 1$ and also $N^{-1}\log \Pi(\Theta_{2/3,\varepsilon}) \to 0$. Therefore, the key inequality (4.15) with $\sigma = 2/3$, $\nu = 1$ holds whenever we choose $s = s_0$ small enough such that

$$-\log 2s_0 > 2\kappa^{-1}\left[2 + \frac{\rho}{2}\left(s_0 - \frac{2}{3} - \varepsilon\right)\right] + 2s_0^2 + \frac{1}{2}.$$

The rest of the detailed derivations follow the same pattern as in the proof of theorem 3.3 and are left to the reader, including verification of (4.11) via an application of theorem 2.5.7 in [61]. In particular, the proof of part **(iii)** follows the same arguments (suppressing the N^{-b} scaling everywhere) as in theorem 3.3.

(f) Proofs for §3*b*

In this section, we prove the results of §3*b* which detail the consequences of the general theorems 3.2 and 3.3 for practical MCMC algorithms.

(i) Proofs for pCN

Theorem 3.6 is proved by verifying the assumption 3.1 for suitable choices of η and L, and for $c_0 = \kappa/2 > 0$.

Lemma 4.9. *Let $\mathcal{P}_N$ denote the transition kernel of pCN from (3.7) with parameter $\beta > 0$.*

(*i*) *Suppose $\Pi = \mathcal{N}(0, I_D/D)$ as in theorem 3.2, and let $Q, \eta > 0$. Then for all $\beta \leq \min\{1/2, \eta/4Q, \eta^2/64\}$ and all $D \geq 1$, we have (with P_0^N-probability 1)*

$$\sup_{\theta \in B_Q} \mathcal{P}_N\left(\theta, \left\{\vartheta : ||\theta - \vartheta||_{\mathbb{R}^D} \geq \frac{\eta}{2}\right\}\right) \leq \mathrm{e}^{-D/2}.$$

(*ii*) *Suppose* $\Pi = \mathcal{N}(0, \Sigma_\alpha)$ *as in theorem 3.3, and let* $Q, \eta > 0$. *There exists some* $c > 0$ *such that for all* $\beta \leq \min\{1/2, \eta/(4Q), c\eta^2/D\}$ *and all* $D \geq 1$, *we have (with* P_0^N*-probability 1)*

$$\sup_{\theta \in B_Q} \mathcal{P}_N\left(\theta, \left\{\vartheta : ||\theta - \vartheta||_{\mathbb{R}^D} \geq \frac{\eta}{2}\right\}\right) \leq \mathrm{e}^{-D/2}.$$

Proof of lemma 4.9. We begin with the proof of part **(ii)**. Let $||\vartheta_k||_{\mathbb{R}^D} \leq Q$. Then using the definition of pCN and that $|\sqrt{1-\beta} - 1| \leq \beta$ for any $\beta \in [0,1]$ (Taylor expanding $\sqrt{\cdot}$ around 1), we obtain that for any $\beta \leq \min\{1/2, \eta/4Q\}$,

$$\begin{aligned}
&\Pr\left(||\vartheta_{k+1} - \vartheta_k||_{\mathbb{R}^D} \geq \frac{\eta}{2}\right) \leq \Pr\left(||p_{k+1} - \vartheta_k||_{\mathbb{R}^D} \geq \frac{\eta}{2}\right) \\
&\leq \Pr\left(||(\sqrt{1-\beta} - 1)\vartheta_k||_{\mathbb{R}^D} + \sqrt{\beta}||\xi_k||_{\mathbb{R}^D} \geq \frac{\eta}{2}\right) \\
&\leq \Pr\left(||\xi_k||_{\mathbb{R}^D} \geq \frac{(\eta/2 - \beta Q)}{\sqrt{\beta}}\right) \\
&\leq \Pr\left(||\xi_k||_{\mathbb{R}^D} \geq \frac{\eta}{4\sqrt{\beta}}\right) \\
&= \Pr\left(||\xi_k||_{\mathbb{R}^D} - \mathbb{E}||\xi_k||_{\mathbb{R}^D} \geq \frac{\eta}{4\sqrt{\beta}} - \mathbb{E}||\xi_k||_{\mathbb{R}^D}\right).
\end{aligned}$$

The variables ξ_k are equal in law to a vector with components $(i^{-\alpha/d} g_i : i \leq D)$ for g_i iid $N(0,1)$ and hence $\mathbb{E}||\xi_k||_{\mathbb{R}^D} \leq (\mathbb{E}||\xi_k||_{\mathbb{R}^D}^2)^{1/2} \leq C(\alpha, d) < \infty$ for $\alpha > d/2$. Then, for $\beta \leq c\eta^2/D$ with some sufficiently small $c > 0$ (noting that then also $\beta \leq c\eta^2$), it holds that

$$\begin{aligned}
\Pr\left(||\vartheta_{k+1} - \vartheta_k||_{\mathbb{R}^D} \geq \frac{\eta}{2}\right) &\leq \Pr\left(||\xi_k||_{\mathbb{R}^D} - \mathbb{E}||\xi_k||_{\mathbb{R}^D} \geq \frac{\eta}{8\sqrt{\beta}}\right) \\
&\leq \exp\left(-\frac{\eta^2}{64\beta}\right) \leq \exp\left(\frac{-D}{2}\right), \qquad (4.31)
\end{aligned}$$

using, e.g. theorem 2.5.8 in [61] (and representing the $||\cdot||_{\mathbb{R}^D}$-norm by duality as a supremum). This completes the proof of part **(ii)**.

The proof of part **(i)** is similar, albeit simpler, whence we leave some details to the reader. Arguing similarly as before, we obtain that for any $\beta \leq \min\{1/2, \eta/64Q\}$,

$$\Pr\left(||\vartheta_{k+1} - \vartheta_k||_{\mathbb{R}^D} \geq \frac{\eta}{2}\right) \leq \Pr\left(||\xi_k||_{\mathbb{R}^D} \geq \frac{(\eta/2 - \beta L)}{\sqrt{\beta}}\right) \leq \Pr\left(||g_k||_{\mathbb{R}^D} \geq \frac{\eta\sqrt{D}}{4\sqrt{\beta}}\right),$$

where g_k is a $\mathcal{N}(0, I_D)$ random variable. The latter probability is bounded by a standard deviation inequality for Gaussians, see, e.g. theorem 2.5.7 in [61]. Indeed, noting that $\mathbb{E}||\xi_k||_{\mathbb{R}^D} \leq (\mathbb{E}[||\xi_k||_{\mathbb{R}^D}^2])^{1/2} = \sqrt{D}$, and that the one-dimensional variances satisfy $\mathbb{E}\langle g_k, v\rangle^2 = ||v||_{\mathbb{R}^D}^2 = 1$ for any $||v||_{\mathbb{R}^D} = 1$, we obtain

$$\begin{aligned}
\Pr(||g_k||_{\mathbb{R}^D} \geq \frac{\eta\sqrt{D}}{4\sqrt{\beta}}) &\leq \Pr\left(\left|||\xi_k||_{\mathbb{R}^D} - \mathbb{E}||\xi_k||_{\mathbb{R}^D}\right| \geq \sqrt{D}\left(\frac{\eta}{4\sqrt{\beta}} - 1\right)\right) \\
&\leq \exp\left(-\frac{D}{2}\left(\frac{\eta}{4\sqrt{\beta}} - 1\right)^2\right) \leq \exp\left(-\frac{D}{2}\right).
\end{aligned}$$

■

Proof of theorem 3.6. We begin with part **(ii)**. Let s_b be as in theorem 3.3 and set $\eta = \eta_N = s_b N^{-b}/2$ as well as $Q = Q_N C\sqrt{N}$, where C is as in theorem 3.3. With those choices, lemma 4.9 **(ii)** implies

that assumption 3.1 is fulfilled with $c_0 = \kappa/2$, so long as β satisfies

$$\beta \leq \min\left\{\frac{1}{2}, \frac{s_b N^{-b}}{8C\sqrt{N}}, \frac{cs_b^2 N^{-2b}}{4D}\right\} \lesssim N^{-2b} D^{-1} \simeq N^{-1-2b}.$$

Hence, the desired result immediately follows from an application of theorem 3.3 **(iv)**.

Part **(i)** of theorem 3.6 similarly follows from verifying assumption 3.1 with $s \in (0, 1/3)$, Q from theorem 3.2, $\eta = s/2$ and for small enough $\beta < c_1$ (with c_1 determined by lemma 4.9 **(i)**), and subsequently applying theorem 3.2 **(iv)**. ■

(ii) Proofs for MALA

Theorem 3.7 is proved by verifying the hypotheses of theorems 3.2 and 3.3, respectively. A key difference between pCN and MALA is that the proposal kernels for MALA, not just its acceptance probabilities, depend on the data $Z^{(N)}$ itself. Again, we begin by examining part **(ii)** which regards $N(0, \Sigma_\alpha)$ priors.

Proof of theorem 3.7, part (ii). We begin by deriving a bound for the gradient $\nabla \log \pi(\cdot|Z^{(N)})$. For Lebesgue-a.e. $\theta \in \mathbb{R}^D$, recalling that $\mathrm{vol}(\mathcal{X}) = 1$, we have that

$$\mathrm{E}_0^N[\nabla \ell_N(\theta)] = -\frac{N}{2} w'(||\theta||) \frac{\theta}{||\theta||} ||g||_{L^2}^2$$

and

$$\begin{aligned}\nabla \ell_N(\theta) &= \frac{1}{2}\sum_{i=1}^N \left(\varepsilon_i - \sqrt{w(||\theta||)} g(X_i)\right) \frac{w'(||\theta||)}{2\sqrt{w(||\theta||)}} \frac{\theta}{||\theta||} g(X_i), \\ &= \frac{w'(||\theta||)}{4\sqrt{w(||\theta||)}} \frac{\theta}{||\theta||} \sum_{i=1}^N \varepsilon_i g(X_i) - \frac{w'(||\theta||)}{4} \frac{\theta}{||\theta||} \sum_{i=1}^N g^2(X_i).\end{aligned}$$

For any $r \in (0, t/2) \cup (t/2, t) \cup (t, L) \cup (L, \infty)$, recalling the choices for T, t, ρ in (4.26) we see that

$$\begin{aligned}\frac{w'(r)}{\sqrt{w(r)}} &= \frac{8Tr}{2Tr} 1_{(0,t/2)}(r) + \frac{T}{\sqrt{w(r)}} 1_{(t/2,t)}(r) + \frac{\rho}{\sqrt{w(r)}} 1_{(t,L)}(r), \\ &\lesssim 1 + N^b + 1, \end{aligned} \tag{4.32}$$

where, to bound the second and third term, we used that $\sqrt{w(r)} \geq Tt = t_b T_b > 0$ is bounded away from zero uniformly in N on $(t/2, \infty)$. Similarly, we have

$$||w'||_\infty \leq \frac{Tt}{2} + T + \rho \lesssim N^b.$$

Combining the above and using Chebyshev's inequality, it follows that

$$\begin{aligned}\sup_{\theta \in \mathbb{R}^D} ||\nabla \ell_N(\theta)||_{\mathbb{R}^D} &\lesssim N^b \left(\left| \sum_{i=1}^N \varepsilon_i g(X_i) \right| + \sum_{i=1}^N g^2(X_i) \right) \\ &\leq N^b \left(||g||_\infty \left| \sum_{i=1}^N \varepsilon_i \right| + \sum_{i=1}^N (g^2(X_i) - ||g||_{L^2}^2) + N||g||_{L^2}^2 \right) \\ &\leq N^b (\mathcal{O}_P(\sqrt{N}) + \mathcal{O}(N)) \\ &= \mathcal{O}(N^{1+b}) + \mathcal{O}(N^{1+b}).\end{aligned}$$

Thus, the event

$$A := \{\sup_{\theta \in \mathbb{R}^D} ||\nabla \ell_N(\theta)||_{\mathbb{R}^D} \leq C' N^{1+b}\},$$

for some large enough $C' > 0$, has probability $P_0^N(A) \to 1$ as $N \to \infty$. We also verify that

$$\nabla \log \pi(\theta) = -\frac{1}{2}\nabla \theta^T \Sigma_\alpha^{-1}\theta = -\Sigma_\alpha^{-1}\theta, \tag{4.33}$$

so that with $Q = Q_N = C\sqrt{N}$ (for C as in theorem 3.3) and recalling that $\Sigma_\alpha = \mathrm{diag}(1, \ldots, \mathrm{D}^{-2\alpha})$, we obtain

$$\sup_{||\theta|| \le Q} ||\nabla \log \pi(\theta)||_{\mathbb{R}^D} = \sup_{||\theta|| \le Q} ||\Sigma_\alpha^{-1}\theta||_{\mathbb{R}^D} \lesssim D^{2\alpha}\sqrt{N} \simeq N^{2\alpha+1}.$$

Now, let s_b also be as in theorem 3.3 and set $\eta = \eta_N = (1/2)s_b N^{-b}$ (note that this is a permissible choice in theorem 3.3). Furthermore, for a small enough constant $c > 0$, let $\gamma \le cN^{-1-2\alpha-b}$. Then since $\alpha > b$, we also have that

$$\gamma \lesssim \min\{N^{-1-2\alpha-b}, N^{-1-2b}, N^{-1/2-b}\}. \tag{4.34}$$

Hence, on the event A and whenever $||\theta||_{\mathbb{R}^D} \le Q$,

$$\gamma ||\nabla \log \pi(\vartheta_k | Z^{(N)})||_{\mathbb{R}^D} \lesssim \gamma(N^{1+b} + N^{1+2\alpha}) \lesssim \eta.$$

Using this, (4.34) and choosing $c > 0$ small enough, conditional on the event A the probability $\Pr(\cdot)$ under the Markov chain satisfies

$$\begin{aligned}
\Pr\left(||p_{k+1} - \vartheta_k|| \ge \frac{\eta}{2}\right) &\le \Pr\left(\gamma ||\nabla \log \pi(\vartheta_k | Z^{(N)})||_{\mathbb{R}^D} \ge \frac{\eta}{4}\right) + \Pr\left(\sqrt{2\gamma}||\xi_{k+1}||_{\mathbb{R}^D} \ge \frac{\eta}{4}\right) \\
&\le \Pr\left(||\xi_{k+1}||_{\mathbb{R}^D} \ge \frac{\eta}{4\sqrt{2\gamma}}\right) \\
&\le \Pr\left(||\xi_{k+1}||_{\mathbb{R}^D} - \mathbb{E}||\xi_{k+1}||_{\mathbb{R}^D} \ge \sqrt{N}\right) \le \exp\left(-\frac{N}{2}\right),
\end{aligned}$$

where the last inequality is proved as in (4.31) above, using theorem 2.5.8 in [61]. Thus, assumption 3.1 is satisfied with $c_0 = 1$ and the proof is complete. ■

Proof of theorem 3.7, part (i). The proof of part (i) proceeds along the same lines, except that (4.32) and (4.33) are replaced with the bound

$$\left|\left|\frac{w'}{\sqrt{w}}\right|\right| s_\infty + ||w'||_\infty < C,$$

for some constant C independent of N, as well as the bound

$$\nabla \log \pi(\theta) = -\frac{D}{2}\nabla ||\theta||^2 = -D\theta, \quad \sup_{||\theta|| \le Q} ||\nabla \log \pi(\theta)||_{\mathbb{R}^D} \simeq NQ.$$

Then letting $s \in (0, 1/3)$ and $Q > 0$ be as in theorem 3.2, and fixing an arbitrary $\eta \in (0, s/2)$, the above implies that for sufficiently small constant $c > 0$ and for any $\gamma \le c/N$, it holds that

$$\begin{aligned}
\Pr\left(||p_{k+1} - \vartheta_k|| \ge \frac{\eta}{2}\right) &\le \Pr\left(\gamma ||\nabla \log \pi(\vartheta_k | Z^{(N)})||_{\mathbb{R}^D} \ge \frac{\eta}{4}\right) + \Pr\left(\sqrt{2\gamma}\xi_{k+1} \ge \frac{\eta}{4}\right) \\
&\le \Pr\left(\xi_{k+1} \ge \frac{\eta}{4\sqrt{2\gamma}}\right) \\
&\le \Pr\left(\xi_{k+1} \ge \frac{\eta\sqrt{\kappa D}}{4\sqrt{2c}}\right).
\end{aligned}$$

Thus, choosing $c > 0$ small enough and arguing exactly as in the last step of the proof of theorem 3.6, part **(i)**, assumption 3.1 is satisfied with $c_0 = 1$ and the proof is complete. ■

Data accessibility. This article has no additional data.

Authors' contributions. R.N.: conceptualization, writing—original draft; A.S.B.: conceptualization, writing—original draft; A.M.: conceptualization, writing—original draft; S.W.: conceptualization, writing—original draft.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. RN was supported by the EPSRC programme grant on Mathematics of Deep Larning, project: EP/V026259.

Acknowledgements. R.N. would like to thank the Forschungsinstitut für Mathematik (FIM) at ETH Zürich for their hospitality during a sabbatical visit in spring 2022 where this research was initiated.

Appendix A. Proofs of §2

Proof of corollary 2.4. We fix $K=1$ and place ourselves under the event of proposition 2.3, and we denote $s=s(\lambda)$ and $t=t(\lambda)$. We can decompose, since $\Pi[\mathcal{S}_s|Y]=1-\Pi[\mathcal{T}_s|Y]$:

$$\Pi(\mathcal{T}_s|Y)=\left[1+\frac{\Pi(\mathcal{S}_s|Y)}{\Pi(\mathcal{T}_s|Y)}\right]^{-1}.$$

Moreover, $\Pi(\mathcal{S}_s|Y)/\Pi(\mathcal{T}_s|Y)=\Pi(\mathcal{S}_s|Y)/[\Pi(\mathcal{T}_t|Y)+\Pi(\mathcal{W}_{s,t}|Y)]\leq\Pi(\mathcal{S}_s|Y)/\Pi(\mathcal{T}_t|Y)$. Using proposition 2.3, for $n\geq n_0(\lambda,Y)$ we have $\Pi(\mathcal{S}_s|Y)/\Pi(\mathcal{T}_s|Y)\leq\exp\{-n\}$. Therefore, $\Pi[\mathcal{T}_s|Y]\geq(1+\exp\{-n\})^{-1}$, which ends the proof. ∎

Proof. The rest of this section is devoted to proving proposition 2.3. We use a uniform bound on the injective norm of Gaussian tensors:

Lemma A.1. *For all $p\geq 3$ there exists a constant C_p, such that:*

$$\limsup_{n\to\infty}\left\{n^{-1/2}\max_{x\in\mathbb{S}^{n-1}}|\langle x^{\otimes p},Z\rangle|\right\}\leq C_p,\quad \textit{almost surely}. \tag{A 1}$$

This lemma is a very crude version of much finer results: in particular the exact value of the constant μ_p such that (w.h.p.) $\max_{x\in\mathbb{S}^{n-1}}|\langle x^{\otimes p},Z\rangle|=\sqrt{n}\mu_p(1+\mathcal{O}_n(1))$ has been first computed non-rigorously in [62], and proven in full generality in [63] (see also discussions in [33,34]). In the rest of this proof, we assume to have conditioned on equation (A 1). For any $0\leq s<t\leq 1$, we have for $n\geq n_0(Y)$:

$$\begin{aligned}\frac{\Pi(\mathcal{S}_s|Y)}{\Pi(\mathcal{T}_t|Y)}&=\frac{\int_{\mathcal{S}_s}\exp(\ell_Y(x))\,\mathrm{d}\Pi(x)}{\int_{\mathcal{T}_t}\exp(\ell_Y(x))\,\mathrm{d}\Pi(x)}\\&\leq \mathrm{e}^{n\lambda C_p}\frac{\int_{\mathcal{S}_s}\exp((n/2)\lambda^2\langle x,x_0\rangle^p)\,\mathrm{d}\Pi(x)}{\int_{\mathcal{T}_t}\exp((n/2)\lambda^2\langle x,x_0\rangle^p)\,\mathrm{d}\Pi(x)},\\&\leq\exp\left(n\lambda C_p+\frac{n\lambda^2}{2}[s^p-t^p]\right)\frac{\Pi(\mathcal{S}_s)}{\Pi(\mathcal{T}_t)}.\end{aligned} \tag{A 2}$$

We upper bound $\Pi(\mathcal{S}_s)\leq\Pi(\mathbb{S}^{n-1})=1$. To lower bound $\Pi(\mathcal{T}_t)$, we use the elementary fact (which is easy to prove using spherical coordinates):

$$\Pi(\mathcal{T}_t)=c_pI_{(1-t)/2}\left[\frac{(n-1)}{2},\frac{(n-1)}{2}\right], \tag{A 3}$$

in which $I_x(a,b)=\int_0^x u^{a-1}(1-u)^{b-1}\,\mathrm{d}u/\int_0^1 u^{a-1}(1-u)^{b-1}\,\mathrm{d}u$ is the incomplete beta function, and $c_p=1$ for odd p and $c_p=2$ for even p. It is then elementary analysis (cf. e.g. [34]) that

$$\lim_{n\to\infty}\frac{1}{n}\log\Pi(\mathcal{T}_t)=\frac{1}{2}\log(1-t^2), \tag{A 4}$$

uniformly in $t\in[0,1)$. Coming back to equation (A 2), this implies that we have, for any $s<t<1$:

$$\limsup_{n\to\infty}\frac{1}{n}\log\frac{\Pi(\mathcal{S}_s|Y)}{\Pi(\mathcal{T}_t|Y)}\leq\lambda C_p+\frac{\lambda^2}{2}[s^p-t^p]-\frac{1}{2}\log(1-t^2). \tag{A 5}$$

Let $K > 0$. It is then elementary to see that it is possible to construct $0 \leq s(\lambda) < t(\lambda) < 1$ with $\lim_{\lambda\to\infty}\{s(\lambda), t(\lambda)\} = 1$, and such that the right-hand side of equation (A 5) becomes smaller than $-K$ as $\lambda \to \infty$. ■

Appendix B. Small ball estimates for isotropic Gaussians

Let $\Pi = \mathcal{N}(0, \mathrm{I}_D/D)$. In this section, we prove equation (4.30), more precisely we show:

Lemma B.1. *Let $a \in (0,1)$. Then for all $D \geq D_0(a)$ large enough, one has for all $z \in (0, 1-a)$:*

$$-\frac{1}{D}\log \Pi(||\theta||_2 \leq z) \geq \frac{1}{2}\left(\frac{z^2}{2} - \log z - \frac{1}{2}\right). \tag{B 1}$$

Proof of lemma B.1. Let $f(x) = -x^2/2 + \log x + 1/2$, so that f reaches its maximum in $x = 1$, with $f(1) = 0$. By decomposition into spherical coordinates and isotropy of the Gaussian measure, one has directly:

$$\Pi(||\theta||_2 \leq z) = \frac{\mathrm{vol}(\mathbb{S}^{D-1})}{(2\pi/D)^{D/2}} \int_0^z \mathrm{d}r\, \mathrm{e}^{-(Dr^2/2)+(D-1)\log r}. \tag{B 2}$$

Recall that $\mathrm{vol}(\mathbb{S}^{D-1}) = 2\pi^{D/2}/\Gamma(D/2)$, so one reaches easily:

$$c_D = \frac{1}{D}\log\frac{\mathrm{vol}(\mathbb{S}^{D-1})}{(2\pi/D)^{D/2}} - \frac{1}{2} = \frac{\log D}{2D} + \mathcal{O}\left(\frac{1}{D}\right). \tag{B 3}$$

In particular, one has for all D large enough (not depending on z):

$$\frac{1}{D}\log \Pi(||\theta||_2 \leq z) \leq \frac{1}{D}\log \int_0^z \mathrm{d}r\, \mathrm{e}^{-(r^2/2)+(D-1)f(r)} + c_D. \tag{B 4}$$

Since f is increasing on $(0,1)$, we have for large enough D:

$$\frac{1}{D}\log \Pi(||\theta||_2 \leq z) \leq \left(1 - \frac{1}{D}\right) f(z) + c_D + \frac{1}{D}\log \int_0^\infty \mathrm{d}r\, \mathrm{e}^{-r^2/2}, \tag{B 5}$$

$$\leq \left(1 - \frac{1}{D}\right) f(z) + \frac{\log D}{D}. \tag{B 6}$$

Since $f(1-a) < 0$, let $D \geq D_0(a)$ large enough such that $f(1-a) \leq -2\log D/(D-2)$. Then for all $z \leq 1-a$, one has $f(z) \leq -2\log D/(D-2)$. Plugging it in the inequality above, we reach that for all $z \in (0, 1-a)$:

$$\frac{1}{D}\log \Pi(||\theta||_2 \leq z) \leq \frac{1}{2} f(z). \tag{B 7}$$

■

References

1. Anderson PW. 1989 Spin glass VI: spin glass as cornucopia. *Phys. Today* **42**, 9.
2. Mézard M, Montanari A. 2009 *Information, physics, and computation*. Oxford, UK: Oxford University Press.
3. Zdeborová L, Krzakala F. 2016 Statistical physics of inference: thresholds and algorithms. *Adv. Phys.* **65**, 453–552.
4. Ben Arous G, Gheissari R, Jagannath A. 2020 Algorithmic thresholds for tensor PCA. *Ann. Probab.* **48**, 2052–2087. (doi:10.1214/19-AOP1415)
5. Ben Arous G, Wein AS, Zadik I. 2020 Free energy wells and overlap gap property in sparse PCA. In *Conf. on Learning Theory, Graz, Austria, 9–12 July 2020*, pp. 479–482. PMLR.
6. Gibbs JW. 1873 A method of geometrical representation of the thermodynamic properties of substances by means of surfaces. *Trans. Conn. Acad. Arts Sci.* **2**, 382–404.

7. Bandeira AS, Alaoui AE, Hopkins SB, Schramm T, Wein AS, Zadik I. 2022 The Franz-Parisi criterion and computational trade-offs in high dimensional statistics. In *Advances in Neural Information Processing Systems* (eds AH Oh, A Agarwal, D Belgrave, K Cho). See https://openreview.net/forum?id=mzze3bubjk.
8. Jerrum M. 2003 *Counting, sampling and integrating: algorithms and complexity*. Lectures in Mathematics ETH Zürich. Basel, Switzerland: Birkhäuser Verlag.
9. Kunisky D, Wein AS, Bandeira AS. 2022 Notes on computational hardness of hypothesis testing: predictions using the low-degree likelihood ratio. In *ISAAC Congress (International Society for Analysis, its Applications and Computation), Aveiro, Portugal, 29 July–2 Aug 2019*, pp. 1–50. Springer.
10. Dalalyan AS. 2017 Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **79**, 651–676. (doi:10.1111/rssb.12183)
11. Durmus A, Moulines E. 2019 High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli* **25**, 2854–2882. (doi:10.3150/18-BEJ1073)
12. Nickl R, Wang S. 2020 On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms. *J. Eur. Math. Soc.*
13. Bohr J, Nickl R. 2021 On log-concave approximations of high-dimensional posterior measures and stability properties in non-linear inverse problems. (http://arxiv.org/abs/2105.07835)
14. Nickl R. 2022 Bayesian non-linear statistical inverse problems. ETH Zurich Lecture Notes.
15. Altmeyer R. 2022 Polynomial time guarantees for sampling based posterior inference in high-dimensional generalised linear models. (http://arxiv.org/abs/2208.13296)
16. Nickl R. 2020 Bernstein–von Mises theorems for statistical inverse problems I: Schrödinger equation. *J. Eur. Math. Soc.* **22**, 2697–2750. (doi:10.4171/JEMS/975)
17. Monard F, Nickl R, Paternain GP. 2019 Efficient nonparametric Bayesian inference for *X*-ray transforms. *Ann. Stat.* **47**, 1113–1147. (doi:10.1214/18-AOS1708)
18. Monard F, Nickl R, Paternain GP. 2021 Statistical guarantees for Bayesian uncertainty quantification in nonlinear inverse problems with Gaussian process priors. *Ann. Stat.* **49**, 3255–3298. (doi:10.1214/21-AOS2082)
19. Monard F, Nickl R, Paternain GP. 2021 Consistent inversion of noisy non-Abelian X-ray transforms. *Commun. Pure Appl. Math.* **74**, 1045–1099. (doi:10.1002/cpa.21942)
20. Stuart AM. 2010 Inverse problems: a Bayesian perspective. *Acta Numer.* **19**, 451–559. (doi:10.1017/S0962492910000061)
21. Fearnhead P, Bierkens J, Pollock M, Roberts GO. 2018 Piecewise deterministic Markov processes for continuous-time Monte Carlo. *Stat. Sci.* **33**, 386–412. (doi:10.1214/18-STS648)
22. Bouchard-Côté A, Vollmer SJ, Doucet A. 2018 The bouncy particle sampler: a nonreversible rejection-free Markov chain Monte Carlo method. *J. Am. Stat. Assoc.* **113**, 855–867.
23. Bierkens J, Grazzi S, Kamatani K, Roberts G. 2020 The boomerang sampler. In *Int. Conf. on Machine Learning, Online, 12–18 July 2020*, pp. 908–918. PMLR.
24. Wu C, Robert CP. 2020 Coordinate sampler: a non-reversible Gibbs-like MCMC sampler. *Stat. Comput.* **30**, 721–730. (doi:10.1007/s11222-019-09913-w)
25. Scalliet C, Guiselin B, Berthier L. 2022 Thirty milliseconds in the life of a supercooled liquid. *Phys. Rev. X***12**, 041028. (doi:10.1103/PhysRevX.12.041028)
26. Grigera TS, Parisi G. 2001 Fast Monte Carlo algorithm for supercooled soft spheres. *Phys. Rev. E* **63**, 045102. (doi:10.1103/PhysRevE.63.045102)
27. Jerrum M. 1992 Large cliques elude the Metropolis process. *Random Struct. Algorithms* **3**, 347–359. (doi:10.1002/rsa.3240030402)
28. Gamarnik D, Zadik I. 2019 The landscape of the planted clique problem: dense subgraphs and the overlap gap property. (http://arxiv.org/abs/1904.07174).
29. Angelini MC, Fachin P, de Feo S. 2021 Mismatching as a tool to enhance algorithmic performances of Monte Carlo methods for the planted clique model. *J. Stat. Mech: Theory Exp.* **2021**, 113406. (doi:10.1088/1742-5468/ac3657)
30. Chen Z, Mossel E, Zadik I. 2023 Almost-linear planted cliques elude the metropolis process. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), Florence, Italy, 22–25 January 2023*, pp. 4504–4539. Philadelphia, PA: SIAM. (doi:10.1137/1.9781611977554.ch171)
31. Hukushima K, Nemoto K. 1996 Exchange Monte Carlo method and application to spin glass simulations. *J. Phys. Soc. Jpn.* **65**, 1604–1608. (doi:10.1143/JPSJ.65.1604)

32. Angelini MC. 2018 Parallel tempering for the planted clique problem. *J. Stat. Mech: Theory Exp.* **2018**, 073404.
33. Richard E, Montanari A. 2014 A statistical model for tensor PCA. In *Advances in neural information processing systems 27, Montreal, Canada, 8–13 Dec 2014*. Red Hook, NY: Curran Associates.
34. Perry A, Wein AS, Bandeira AS. 2020 Statistical limits of spiked tensor models. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **56**, 230–264.
35. Lesieur T, Miolane L, Lelarge M, Krzakala F, Zdeborová L. 2017 Statistical and computational phase transitions in spiked tensor estimation. In *2017 IEEE Int. Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017*, pp. 511–515. New York, NY: IEEE.
36. Jagannath A, Lopatto P, Miolane L. 2020 Statistical thresholds for tensor PCA. *Ann. Appl. Probab.* **30**, 1910–1933. (doi:10.1214/19-AAP1547)
37. Baik J, Ben Arous G, Péché S. 2005 Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Appl. Probab.* **33**, 1643–1697.
38. Wein AS, El Alaoui A, Moore C. 2019 The Kikuchi hierarchy and tensor PCA. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS), Baltimore, MA, 9–12 November 2019*, pp. 1446–1468. New York, NY: IEEE.
39. Hopkins SB, Shi J, Steurer D. 2015 Tensor principal component analysis via sum-of-square proofs. In *Conf. on Learning Theory, Paris, France, 3–6 July 2015*, pp. 956–1006. PMLR.
40. Hopkins SB, Schramm T, Shi J, Steurer D. 2016 Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *Proc. of the Forty-Eighth Annual ACM Symposium on Theory of Computing, Cambridge, MA, 19–21 June 2016*, pp. 178–191. New York, NY: ACM.
41. Kim C, Bandeira AS, Goemans MX. 2017 Community detection in hypergraphs, spiked tensor models, and sum-of-squares. In *2017 International Conference on Sampling Theory and Applications (SampTA), Bordeaux, France, 8–12 July 2017*, pp. 124–128. New York, NY: IEEE.
42. Sarao Mannelli S, Biroli G, Cammarota C, Krzakala F, Zdeborová L. 2019 Who is afraid of big bad minima? Analysis of gradient-flow in spiked matrix-tensor models. *Advances in neural information processing systems32, Vancouver, Canada, 8–14 Dec 2019*. Red Hook, NY: Curran Associates.
43. Sarao Mannelli S, Krzakala F, Urbani P, Zdeborová L. 2019 Passed & spurious: descent algorithms and local minima in spiked matrix-tensor models. In *International Conference on Machine Learning*, pp. 4333-4342. PMLR.
44. Biroli G, Cammarota C, Ricci-Tersenghi F. 2020 How to iron out rough landscapes and get optimal performances: averaged gradient descent and its application to tensor PCA. *J. Phys. A: Math. Theor.* **53**, 174003. (doi:10.1088/1751-8121/ab7b1f)
45. Ben Arous G, Gheissari R, Jagannath A. 2020 Bounding flows for spherical spin glass dynamics. *Commun. Math. Phys.* **373**, 1011–1048. (doi:10.1007/s00220-019-03649-4)
46. Ben Arous G, Gheissari R, Jagannath A. 2021 Online stochastic gradient descent on non-convex losses from high-dimensional inference. *J. Mach. Learn. Res.* **22**, 106–1.
47. Rasmussen CE, Williams CKI. 2006 *Gaussian processes for machine learning*. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press.
48. Ghosal S, van der Vaart AW. 2017 *Fundamentals of nonparametric Bayesian inference*. New York, NY: Cambridge University Press.
49. van der Vaart A, van Zanten JH. 2008 Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Stat.* **36**, 1435–1463. (doi:10.1214/009053607000000613)
50. Cotter SL, Roberts GO, Stuart AM, White D. 2013 MCMC methods for functions: modifying old algorithms to make them faster. *Stat. Sci.* **28**, 424–446. (doi:10.1214/13-STS421)
51. Beskos A, Girolami M, Lan S, Farrell PE, Stuart AM. 2017 Geometric MCMC for infinite-dimensional inverse problems. *J. Comput. Phys.* **335**, 327–351. (doi:10.1016/j.jcp.2016.12.041)
52. Hairer M, Stuart AM, Vollmer SJ. 2014 Spectral gaps for a Metropolis-Hastings algorithm in infinite dimensions. *Ann. Appl. Probab.* **24**, 2455–2490. (doi:10.1214/13-AAP982)
53. Hairer M, Mattingly J, Scheutzow M. 2011 Asymptotic coupling and a general form of Harris' theorem with applications to stochastic delay equations. *Probab. Theory Relat. Fields* **149**, 223–259. (doi:10.1007/s00440-009-0250-6)

54. Chewi S, Lu C, Ahn K, Cheng X, Le Gouic T, Rigollet P. 2021 Optimal dimension dependence of the Metropolis-Adjusted Langevin Algorithm. In *Conference on Learning Theory, Boulder, CO, 15–19 August 2021*. PMLR.
55. Roberts GO, Rosenthal JS. 2001 Optimal scaling for various Metropolis-Hastings algorithms. *Stat. Sci.* **16**, 351–367. (doi:10.1214/ss/1015346320)
56. Breyer LA, Piccioni M, Scarlatti S. 2004 Optimal scaling of MaLa for nonlinear regression. *Ann. Appl. Probab.* **14**, 1479–1505. (doi:10.1214/105051604000000369)
57. Mattingly JC, Pillai NS, Stuart AM. 2012 Diffusion limits of the random walk Metropolis algorithm in high dimensions. *Ann. Appl. Probab.* **22**, 881–930. (doi:10.1214/10-AAP754)
58. Franz S, Parisi G. 1995 Recipes for metastable states in spin glasses. *J. Phys. I* **5**, 1401–1415.
59. Edmunds DE, Triebel H. 1996 *Function spaces, entropy numbers, differential operators*. Cambridge Tracts in Mathematics, vol. 120. Cambridge, UK: Cambridge University Press.
60. Li WV, Linde W. 1999 Approximation, metric entropy and small ball estimates for Gaussian measures. *Ann. Probab.* **27**, 1556–1578. (doi:10.1214/aop/1022677459)
61. Giné E, Nickl R. 2016 *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics. New York, NY: Cambridge University Press.
62. Crisanti A, Sommers HJ. 1992 The spherical *p*-spin interaction spin glass model: the statics. *Zeitschrift für Physik B Condensed Matter* **87**, 341–354. (doi:10.1007/BF01309287)
63. Subag E. 2017 The complexity of spherical *p*-spin models – a second moment approach. *Ann. Probab.* **45**, 3385–3450. (doi:10.1214/16-AOP1139)