On the computational complexity of MCMC in high-dimensional non-linear regression models

Richard Nickl

University of Cambridge (UK)

Conference for Elisabeth, Paris, 1st of June 2023
Consider statistical observations arising as random vectors

\[ Y_i = G_\theta(X_i) + \varepsilon_i, \quad \varepsilon_i \sim^{i.i.d.} \mathcal{N}_V(0, I), \]

- The \( X_i \) are covariates drawn iid from some law \( \lambda \) on a \( d \)-dimensional domain \( \mathcal{X} \).
- The \( Y_i \) are ‘response’ variables in a vector space \( V \) of finite \( \text{dim}(V) \), say \( = 1 \).

We write \( Z^{(N)} = (Y_i, X_i)_{i=1}^N \) for the data vector of sample size \( N \in \mathbb{N} \).
Consider statistical observations arising as random vectors

\[ Y_i = G_\theta(X_i) + \varepsilon_i, \quad \varepsilon_i \sim_{i.i.d.} \mathcal{N}_V(0, I), \]

- The \( X_i \) are covariates drawn iid from some law \( \lambda \) on a \( d \)-dimensional domain \( \mathcal{X} \).
- The \( Y_i \) are ‘response’ variables in a vector space \( V \) of finite \( \text{dim}(V) \), say = 1.

We write \( Z^{(N)} = (Y_i, X_i)_{i=1}^N \) for the data vector of sample size \( N \in \mathbb{N} \).

The regression fields

\[ \{G_\theta : \theta \in \Theta\}, \quad G_\theta : \mathcal{X} \to V, \]

are indexed by the high-dimensional parameter

\[ \theta \in \Theta = \mathbb{R}^D \]

arising from the discretisation of a function space in some basis. \textbf{We take asymptotics} \( D, N \to \infty, \text{ possibly } D/N \to \kappa > 0 \).
The dependence $\theta \mapsto G(\theta)$ is non-linear and we have in mind coefficient to solution maps of partial differential equations (PDEs).
The dependence $\theta \mapsto G(\theta)$ is non-linear and we have in mind coefficient to solution maps of partial differential equations (PDEs).

A standard model problem concerns inference on the diffusivity coefficient $f_\theta = e^\theta > 0$ from (unique) solutions $u = u_\theta = G_\theta$ of the elliptic PDE

$$\nabla \cdot (f_\theta \nabla u) = g \text{ in } \mathcal{X},$$

$$u = 0 \text{ on } \partial \mathcal{X}.$$
The dependence $\theta \mapsto \mathcal{G}(\theta)$ is non-linear and we have in mind coefficient to solution maps of partial differential equations (PDEs).

A standard model problem concerns inference on the diffusivity coefficient $f_{\theta} = e^\theta > 0$ from (unique) solutions $u = u_{\theta} = \mathcal{G}_{\theta}$ of the elliptic PDE

$$\nabla \cdot (f_{\theta} \nabla u) = g \text{ in } \mathcal{X},$$

$$u = 0 \text{ on } \partial \mathcal{X}.$$ 

This can be regarded as a ‘steady state’ measurement of the process of diffusion. Inferring $\theta$ is sometimes known as ‘Darcy’s problem’ (Stuart 2010).
• The dependence \( \theta \mapsto G(\theta) \) is non-linear and we have in mind coefficient to solution maps of partial differential equations (PDEs).

• A standard model problem concerns inference on the diffusivity coefficient \( f_\theta = e^\theta > 0 \) from (unique) solutions \( u = u_\theta = G_\theta \) of the elliptic PDE

\[
\nabla \cdot (f_\theta \nabla u) = g \text{ in } \mathcal{X},
\]
\[
u = 0 \text{ on } \partial \mathcal{X}.
\]

This can be regarded as a ‘steady state’ measurement of the process of diffusion. Inferring \( \theta \) is sometimes known as ‘Darcy’s problem’ (Stuart 2010).

• The ‘forward’ map \( \theta \mapsto G_\theta \) can be ‘evaluated’ by numerical PDE methods (finite elements, etc.).
A natural approach (Gauß, Tikhonov) is to minimise a least squares fit

\[ Q_N(\theta) = \sum_{i=1}^{N} |Y_i - G_{\theta}(X_i)|^2 + \lambda \cdot pen(\theta), \ \lambda > 0, \]

over \( \theta \in \mathbb{R}^D \). The penalty term is also called a ‘regulariser’.
A natural approach (Gauß, Tikhonov) is to minimise a least squares fit

\[
Q_N(\theta) = \sum_{i=1}^{N} |Y_i - G_\theta(X_i)|_V^2 + \lambda \cdot \text{pen}(\theta), \quad \lambda > 0,
\]

over \( \theta \in \mathbb{R}^D \). The penalty term is also called a ‘regulariser’.

Classical choices are Sobolev type norms

\[
\text{pen}(\theta) = \|\theta\|_{H^\alpha}^2 \simeq \sum_{j \leq D} j^{2\alpha/d} |\theta_j|^2, \quad \alpha \in \mathbb{N},
\]

or related \( \ell_1 \)-type penalties/TV-type norms.
A natural approach (Gauß, Tikhonov) is to minimise a least squares fit

\[ Q_N(\theta) = \sum_{i=1}^{N} |Y_i - G_\theta(X_i)|^2 + \lambda \cdot \text{pen}(\theta), \quad \lambda > 0, \]

over \( \theta \in \mathbb{R}^D \). The penalty term is also called a ‘regulariser’.

Classical choices are Sobolev type norms

\[ \text{pen}(\theta) = \|\theta\|_{H^\alpha}^2 \approx \sum_{j \leq D} j^{2\alpha/d} |\theta_j|^2, \quad \alpha \in \mathbb{N}, \]

or related \( \ell_1 \)-type penalties/TV-type norms.

As \( G \) is non-linear in \( \theta \), the map \( Q_N \) is not convex on \( \mathbb{R}^D \).
A natural approach (Gauß, Tikhonov) is to minimise a least squares fit

\[ Q_N(\theta) = \sum_{i=1}^{N} |Y_i - G_\theta(X_i)|^2 + \lambda \cdot \text{pen}(\theta), \quad \lambda > 0, \]

over \( \theta \in \mathbb{R}^D \). The penalty term is also called a ‘regulariser’.

Classical choices are Sobolev type norms

\[ \text{pen}(\theta) = \|\theta\|^2_{H^\alpha} \simeq \sum_{j \leq D} j^{2\alpha/d} |\theta_j|^2, \quad \alpha \in \mathbb{N}, \]

or related \( \ell_1 \)-type penalties/TV-type norms.

As \( G \) is non-linear in \( \theta \), the map \( Q_N \) is not convex on \( \mathbb{R}^D \).

The algorithmic runtime = \( \{ \# \text{ required evaluations of } G(\theta) \} \) to compute such optimisers may scale exponentially in dimension \( D \) and sample size \( N \).
Consider a centred Gaussian process \( (X(z) : z \in \mathcal{Z} \subset \mathbb{R}^d) \), with covariance \( K(y, z) = k_\alpha(z - y) \) where

\[
k_\alpha(z) = \int_{\mathbb{R}^d} e^{i \langle z, \xi \rangle} d\mu(\xi), \quad d\mu(\xi) = (1 + |\xi|^2)^{-\alpha} d\xi, \quad \alpha > d/2,
\]

modelling \( \alpha \)-regular stationary random fields \( \theta \) over \( \mathcal{Z} \), with RKHS \( \mathcal{H} = H^\alpha \).
Gaussian processes models for functions

Consider a centred Gaussian process \( (X(z) : z \in \mathcal{Z} \subset \mathbb{R}^d) \), with covariance
\[
K(y, z) = k_\alpha(z - y)
\]
where
\[
k_\alpha(z) = \int_{\mathbb{R}^d} e^{i\langle z, \xi \rangle} d\tilde{\mu}(\xi), \quad d\tilde{\mu}(\xi) = (1 + |\xi|^2)^{-\alpha} d\xi, \quad \alpha > d/2,
\]
modelling \( \alpha \)-regular stationary random fields \( \theta \) over \( \mathcal{Z} \), with RKHS \( \mathcal{H} = H^\alpha \).

- Such ‘Whittle-Matérn’ processes are often discretised by projection onto the first \( D \) eigenfunctions of the Laplacian (\( = \) Karhunen-Loève expansion) or other bases, and as such used as **Bayesian priors** for \( \theta \) and then also \( \mathcal{G}_\theta \) in inverse problems.
Bayesian Inversion with Gaussian priors

The ‘log-likelihood’ of our statistical regression model is

\[ \ell_N(\theta) = \log dP^N_\theta(Z^{(N)}) - \text{const} = -\frac{1}{2} \sum_{i=1}^{N} |Y_i - G_\theta(X_i)|^2_V, \quad \theta \in \mathbb{R}^D, \]

whose evaluation requires only computation of \( G(\theta) \) (forward PDE).
Bayesian Inversion with Gaussian priors

The ‘log-likelihood’ of our statistical regression model is

\[ \ell_N(\theta) = \log dP^N_\theta(Z^{(N)}) - \text{const} = -\frac{1}{2} \sum_{i=1}^N |Y_i - G_\theta(X_i)|_V^2, \quad \theta \in \mathbb{R}^D, \]

whose evaluation requires only computation of \( G(\theta) \) (forward PDE).

Let \( \Pi = \Pi_D \) be the \( D \)-dimensional discretisation of a Gaussian process prior with RKHS \( \mathcal{H} \). The posterior distribution \( \Pi(\cdot | Z^{(N)}) \) given data \( Z^{(N)} \) on \( \mathbb{R}^D \) equals

\[ d\Pi(\theta | Z^{(N)}) \propto e^{\ell_N(\theta)} d\Pi(\theta) \propto e^{\ell_N(\theta)} - \frac{1}{2} \| \theta \|_\mathcal{H}^2, \quad \theta \in \mathbb{R}^D. \]
Bayesian Inversion with Gaussian priors

The ‘log-likelihood’ of our statistical regression model is

$$
\ell_N(\theta) = \log dP_\theta^N(Z^{(N)}) - \text{const} = -\frac{1}{2} \sum_{i=1}^{N} |Y_i - G_\theta(X_i)|^2_V, \quad \theta \in \mathbb{R}^D,
$$

whose evaluation requires only computation of $G(\theta)$ (forward PDE).

Let $\Pi = \Pi_D$ be the $D$-dimensional discretisation of a Gaussian process prior with RKHS $\mathcal{H}$. The posterior distribution $\Pi(\cdot|Z^{(N)})$ given data $Z^{(N)}$ on $\mathbb{R}^D$ equals

$$
d\Pi(\theta|Z^{(N)}) \propto e^{\ell_N(\theta)} d\Pi(\theta) \propto e^{\ell_N(\theta)} - \frac{1}{2} \|\theta\|^2_\mathcal{H}, \quad \theta \in \mathbb{R}^D.
$$

If a Markov chain $(\vartheta_k)$ on $\mathbb{R}^D$ has invariant measure $\Pi(\cdot|Z^{(N)})$, we can approximately compute the posterior mean vector

$$
E^{\Pi}[\theta|Z^{(N)}] = \int_{\mathbb{R}^D} \theta d\Pi(\theta|Z^{(N)})
$$

by ergodic ‘sample’ averages $\frac{1}{K} \sum_{k=1}^{K} \vartheta_k$ accrued along the chain.

**pre-conditioned Crank-Nicolson (pCN) algorithm**

Let $\Pi \sim N(0, \mathcal{K})$ on $\Theta = \mathbb{R}^D$. Fix $\delta > 0$ and initialise $\vartheta_0$. For $k \geq 0$ do:

1. Draw $\xi \sim \Pi$ and calculate the proposal

$$p_{\vartheta_k} = \sqrt{1 - 2\delta \vartheta_k} + \sqrt{2\delta} \xi.$$ 

2. Set

$$\vartheta_{k+1} = \begin{cases} p_{\vartheta_k}, & \text{with probability } 1 \wedge \exp\{\ell_N(p_{\vartheta_k}) - \ell_N(\vartheta_k)\} \\ \vartheta_k, & \text{else.} \end{cases}$$

A standard ‘Metropolis-Hastings’ calculation shows that $\{\vartheta_k\}$ has invariant measure $\Pi(\cdot | Z^{(N)})$. 

Discretised Langevin type algorithm (ULA)

Choose step size $\delta > 0$ and an initialiser $\vartheta_0$. For $\xi_k \sim iid \mathcal{N}(0, I)$ in $\mathbb{R}^D$, do:

$$\vartheta_{k+1} = \vartheta_k + \delta \nabla \log d\Pi(\vartheta_k | Z^{(N)}) + \sqrt{2\delta} \xi_k, \quad k \in \mathbb{N}.$$
Discretised Langevin type algorithm (ULA)

Choose step size $\delta > 0$ and an initialiser $\vartheta_0$. For $\xi_k \sim iid \mathcal{N}(0, I)$ in $\mathbb{R}^D$, do:

$$\vartheta_{k+1} = \vartheta_k + \delta \nabla \log d\Pi(\vartheta_k | Z^{(N)}) + \sqrt{2\delta} \xi_k, \quad k \in \mathbb{N}.$$ 

The discretisation step mis-specifies the invariant measure, but that ‘bias’ decreases as $\delta \to 0$.

As before, one can add a Metropolis-Hastings adjustment (MALA) accept/reject step to obtain the exact posterior distribution as invariant measure.
Is posterior computation by MCMC possible in such non-linear regression problems \textbf{in polynomial run time} in dimension $D$ and sample size (informativeness) $N$?
Is posterior computation by MCMC possible in such non-linear regression problems in polynomial run time in dimension $D$ and sample size (informativeness) $N$?

Let us focus on computation of the $D$-dimensional integral

$$E^n[\theta | Z^{(N)}] = \frac{\int_{\mathbb{R}^D} \theta e^{\ell_N(\theta)} d\Pi(\theta) d\theta}{\int_{\mathbb{R}^D} e^{\ell_N(\theta)} d\Pi(\theta) d\theta}.$$
Is posterior computation by MCMC possible in such non-linear regression problems in polynomial run time in dimension $D$ and sample size (informativeness) $N$?

Let us focus on computation of the $D$-dimensional integral

$$E^\Pi[\theta|Z^{(N)}] = \frac{\int_{\mathbb{R}^D} \theta e^{\ell_N(\theta)} d\Pi(\theta) d\theta}{\int_{\mathbb{R}^D} e^{\ell_N(\theta)} d\Pi(\theta) d\theta}.$$

The worst case deterministic numerical cost for evaluating the integral of a $D$-dimensional 1-Lipschitz function scales as $D^{D/4}$ (Novak & Wozniakowski 2008).
Is posterior computation by MCMC possible in such non-linear regression problems in polynomial run time in dimension $D$ and sample size (informativeness) $N$?

Let us focus on computation of the $D$-dimensional integral

$$E^{\Pi}[\theta|Z^{(N)}] = \frac{\int_{\mathbb{R}^D} \theta e^{\ell_{N}(\theta)} d\Pi(\theta) d\theta}{\int_{\mathbb{R}^D} e^{\ell_{N}(\theta)} d\Pi(\theta) d\theta}.$$ 

The worst case deterministic numerical cost for evaluating the integral of a $D$-dimensional 1-Lipschitz function scales as $D^{D/4}$ (Novak & Wozniakowski 2008).

Randomised (‘Monte Carlo’) algorithms may beat such computational barriers with universal accuracy $1/\sqrt{K}$ after $K$ iterations (central limit theorem).
In the class of strongly globally log-concave target measures, Langevin-type algorithms achieve polynomial mixing time in $D, N$ with high probability for any precision level (in $W_2$-distance).

This applies to linear $G$ and Gaussian priors as the posterior is then log-concave.
In the class of **strongly globally log-concave** target measures, Langevin-type algorithms achieve polynomial mixing time in $D, N$ with high probability for any precision level (in $W_2$-distance).

This applies to **linear $G$** and Gaussian priors as the posterior is then log-concave.

The theory extends to target measures satisfying a **log-Sobolev inequality** (e.g., Vempala & Wibisono (2021)), but the LSI-constants scale *exponentially* in bounded (‘Holley-Stroock’) perturbations.
Hardness of posterior computation for local cold start MCMC

[Joint work with A. Bandeira, A. Maillard, S. Wang, (2023)]
A radial average negative log-likelihood

We consider posteriors arising from $\alpha$-regular Gaussian process priors and expected log-likelihoods $-E_{\theta_0} \ell_N(\theta) = -\frac{N}{2} - \frac{N}{2} w(\|\theta - \theta_0\|)$, with $w$ of the form

![Graph showing the function $w(r)$ with local convexity near 0 and linear growth from $t/2$ onwards.](image)

which is locally convex near 0 and then grows piece-wise linearly from $t/2$ onwards. We consider ‘local’ algorithms initialised in $[t, L]$ where $w$ exhibits linear growth.
Recall that for Whittle-Matern $N(0, \Sigma_{\alpha})$-prior, the posterior on $\Theta = \mathbb{R}^D$ is

$$d\Pi(\theta|Z^{(N)}) \propto \exp (\ell_N(\theta) - \frac{1}{2}\theta^T\Sigma^{-1}_{\alpha}\theta), \quad \theta \in \mathbb{R}^D.$$
Recall that for Whittle-Matern $N(0, \Sigma_\alpha)$-prior, the posterior on $\Theta = \mathbb{R}^D$ is

\[
d\Pi(\theta|Z^{(N)}) \propto \exp \left( \ell_N(\theta) - \frac{1}{2} \theta^T \Sigma_\alpha^{-1} \theta \right), \quad \theta \in \mathbb{R}^D.
\]

Define Euclidean balls centred at the ground truth $\theta_0 \in \mathbb{R}^D$,

\[B_r = \{ \theta \in \mathbb{R}^D : \| \theta - \theta_0 \|_{\mathbb{R}^D} \leq r \}\]

and $D$-dimensional annuli

\[\Theta_{r,\varepsilon} = \{ \theta \in \mathbb{R}^D : \| \theta - \theta_0 \|_{\mathbb{R}^D} \in (r, r + \varepsilon] \} = B_{r+\varepsilon} \setminus B_r.
\]

We consider non-intersecting ‘inner’ and ‘outer’ annuli $\Theta_{s,\eta}$ and $\Theta_{r,\varepsilon}$, with $s < \sigma$. 
Recall that for Whittle-Matern $N(0, \Sigma(\alpha))$-prior, the posterior on $\Theta = \mathbb{R}^D$ is

$$d\Pi(\theta|Z^{(N)}) \propto \exp \left( \ell_N(\theta) - \frac{1}{2} \theta^T \Sigma^{-1} \theta \right), \quad \theta \in \mathbb{R}^D.$$ 

Define Euclidean balls centred at the ground truth $\theta_0 \in \mathbb{R}^D$,

$$B_r = \{ \theta \in \mathbb{R}^D : \|\theta - \theta_0\|_{\mathbb{R}^D} \leq r \}$$

and $D$-dimensional annuli

$$\Theta_{r,\varepsilon} = \{ \theta \in \mathbb{R}^D : \|\theta - \theta_0\|_{\mathbb{R}^D} \in (r, r + \varepsilon] \} = B_{r+\varepsilon} \setminus B_r.$$ 

We consider non-intersecting ‘inner’ and ‘outer’ annuli $\Theta_{s,\eta}$ and $\Theta_{r,\varepsilon}$, with $s < \sigma$. We will assume $\theta_0 = 0$ is the ground truth so that the prior is already centred at the correct value, and the ‘picture’ is centred at the origin.
Consider any Markov chain \((\vartheta_k : k \in \mathbb{N})\) with invariant ‘target’ measure \(\mu\) (e.g., \(\mu = \Pi(\cdot | Z^{(N)})\)) for which the ratio bound
\[
\frac{\mu(\Theta_s, \eta)}{\mu(\Theta_{\sigma, \varepsilon})} \leq e^{-\nu N}
\]
holds for some \(\nu > 0\). For constants \(\eta < \sigma - s\), suppose \(\vartheta_0\) is started in the ‘outer annulus’ \(\Theta_{\sigma, \varepsilon}\), drawn from the \textit{conditional} distribution \(\mu(\cdot | \Theta_{\sigma, \varepsilon})\), and denote by
\[
\tau_s = \inf\{k : \vartheta_k \in \Theta_{s, \eta}\}
\]
the \textit{hitting time} of the Markov chain onto the intermediate annulus \(\Theta_{s, \eta}\). Then
\[
\Pr(\tau_s \leq K) \leq Ke^{-\nu N}, \text{ for all } K > 0.
\]
Because of monotonic, radial growth of \(-E_{\theta_0} \ell_N(\theta)\), with high prob. and for \(\sigma = 1\),

\[
\frac{1}{N} \log \frac{\prod(\Theta_s, \eta | Z^{(N)})}{\prod(\Theta_1, \epsilon | Z^{(N)})} \leq \frac{1}{N} \log \frac{\prod(\Theta_s, \eta)}{\prod(\Theta_1, \epsilon)} + w(1 + \epsilon) - w(s) + o_P(1).
\]
Because of monotonic, radial growth of $-E_{\theta_0} \ell_N(\theta)$, with high prob. and for $\sigma = 1$,

$$\frac{1}{N} \log \frac{\Pi(\Theta_{s,\eta}|Z^{(N)})}{\Pi(\Theta_{1,\varepsilon}|Z^{(N)})} \leq \frac{1}{N} \log \frac{\Pi(\Theta_{s,\eta})}{\Pi(\Theta_{1,\varepsilon})} + w(1 + \varepsilon) - w(s) + o_P(1).$$

In high dimensions a ‘free energy barrier’ can appear (cf. Ben Arous, Wein, Zadik (2022, CPAM) in spin glass models with uniform priors), because the ‘intermediate annulus’ $\Theta_{s,\eta}$ has much smaller Gaussian volume than the outer annulus $\Theta_{1,1+\varepsilon}$. 
Because of monotonic, radial growth of $-E_{\theta_0} \ell_N(\theta)$, with high prob. and for $\sigma = 1$,\[
\frac{1}{N} \log \frac{\Pi(\Theta_{s,\eta} | Z^{(N)})}{\Pi(\Theta_{1,\varepsilon} | Z^{(N)})} \leq \frac{1}{N} \log \frac{\Pi(\Theta_{s,\eta})}{\Pi(\Theta_{1,\varepsilon})} + w(1 + \varepsilon) - w(s) + o_P(1) .
\]

In high dimensions a ‘free energy barrier’ can appear (cf. Ben Arous, Wein, Zadik (2022, CPAM) in spin glass models with uniform priors), because the ‘intermediate annulus’ $\Theta_{s,\eta}$ has much smaller Gaussian volume than the outer annulus $\Theta_{1,1+\varepsilon}$.

For $\alpha$-regular Gaussian process priors with $\eta = o(N^{-b})$, $b = b_{\alpha,d} > 0$, and $D/N \sim \kappa > 0$, this barrier is non-degenerate at the $(1/N) \log$ scale:
\[
\frac{1}{N} \log \frac{\Pi(\Theta_{s,\eta})}{\Pi(\Theta_{1,\varepsilon})} \leq -\nu, \quad \nu > 0, \quad \text{some } \varepsilon > 0 .
\]
Large Deviation Landscape (Franz-Parisi functional)

Because of monotonic, radial growth of $-E_{\theta_0} \ell_N(\theta)$, with high prob. and for $\sigma = 1$,

$$
\frac{1}{N} \log \frac{\Pi(\Theta_s, \eta | Z^{(N)})}{\Pi(\Theta_1, \epsilon | Z^{(N)})} \leq \frac{1}{N} \log \frac{\Pi(\Theta_s, \eta)}{\Pi(\Theta_1, \epsilon)} + w(1 + \epsilon) - w(s) + o_P(1).
$$

In high dimensions a ‘free energy barrier’ can appear (cf. Ben Arous, Wein, Zadik (2022, CPAM) in spin glass models with uniform priors), because the ‘intermediate annulus’ $\Theta_s, \eta$ has much smaller Gaussian volume than the outer annulus $\Theta_1, 1+\epsilon$.

For $\alpha$-regular Gaussian process priors with $\eta = o(N^{-b})$, $b = b_{\alpha,d} > 0$, and $D/N \sim \kappa > 0$, this barrier is non-degenerate at the $(1/N) \log$ scale:

$$
\frac{1}{N} \log \frac{\Pi(\Theta_s, \eta)}{\Pi(\Theta_1, \epsilon)} \leq -\nu, \quad \nu > 0, \quad \text{some } \epsilon > 0.
$$

We show that this does not prevent the posterior to charge all its mass inside of $B_s –$ the barrier lies outside of the region where the posterior concentrates.
Let $P_N(\theta, A)$ denote a sequence of kernels describing the transition dynamics from $\theta \in \mathbb{R}^D$ into $A \subset \mathbb{R}^D$ of a Markov chain $(\vartheta_k)$.

**Condition (A)**

i) $P_N(\cdot, \cdot)$ has invariant distribution $\Pi(\cdot | Z^{(N)})$.

ii) For some fixed $c_0, L > 0, \eta = \eta_N > 0$, with high prob.,

$$\sup_{\theta \in B_L} P_N(\theta, \{\vartheta : \|\theta - \vartheta\|_{\mathbb{R}^D} \geq \eta/2\}) \leq e^{-c_0 N}, \quad N \geq 1.$$ 

The second condition means that the MCMC moves ‘locally’, with large steps being unlikely to occur. This condition can be verified for pCN and MALA with natural parameter choices.
Theorem

Let $\Pi(\cdot|Z^{(N)})$ arise from prior $N(0, \Sigma_\alpha)$, $b = \alpha/d - 1/2 > 0$ and $D/N \simeq \kappa > 0$. There exists $s_b \in (0, 1/2)$ s.t.:
A general hitting time lower bound

**Theorem**

Let \( \Pi(\cdot|Z^{(N)}) \) arise from prior \( N(0, \Sigma_{\alpha}) \), \( b = \alpha/d - 1/2 > 0 \) and \( D/N \simeq \kappa > 0 \). There exists \( s_b \in (0, 1/2) \) s.t.:

i) The average log-likelihood is unimodal with mode 0, locally log-concave near 0, radially symmetric, Lipschitz continuous and decreasing in \( \|\theta\|_{\mathbb{R}^D} \) on \( \mathbb{R}^D \).

ii) For any \( r > 0 \) fixed and 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 \( \Pi(B_s|Z^{(N)}) \) as \( N \to \infty \to \to 1 \) in probability.

iv) There exist \( \varepsilon, C > 0 \) s.t. for any Markov kernels \( P_N \) on \( \mathbb{R}^D \) and associated chains \( (\vartheta_k) \) satisfying Condition (A) for \( \eta_N \in (0, s_b N - b) \), we can find an initialiser \( \vartheta_0 \in \Theta_N - b, \varepsilon \) s.t. w.h.p. the hitting time \( \tau_{B_s} \) for \( \vartheta_k \) to reach \( B_s \) is

\[
\tau_{B_s} \geq \exp\left(\min\{c_0, 1\} D/2\right)
\]

In fact we can take \( \vartheta_0 \sim \mu | \Theta_N - b, \varepsilon \) with \( \mu = \Pi(\cdot|Z^{(N)}) \).
A general hitting time lower bound

Theorem

Let $\Pi(\cdot | Z^{(N)})$ arise from prior $N(0, \Sigma_{\alpha})$, $b = \alpha/d - 1/2 > 0$ and $D/N \simeq \kappa > 0$. There exists $s_b \in (0, 1/2)$ s.t.:

i) The average log-likelihood is unimodal with mode 0, locally log-concave near 0, radially symmetric, Lipschitz continuous and decreasing in $\|\theta\|_{\mathbb{R}^D}$ on $\mathbb{R}^D$.

ii) For any $r > 0$ fixed and 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 $s_b(N) = s_bN^{-b}$, we have $\Pi(B_s|Z^{(N)}) \Rightarrow 1$ in probability.

iv) There exist $\epsilon, C > 0$ s.t. for any Markov kernels $P_N$ on $\mathbb{R}^D$ and associated chains $(\vartheta_k)$ satisfying Condition (A) for $\eta_N \in (0, s_b(N)-b)$, we can find an initialiser $\vartheta_0 \in \Theta_N - b, \epsilon_N - b$ s.t. w.h.p. the hitting time $\tau_{B_s}$ for $\vartheta_k$ to reach $B_s$ is

$$\tau_{B_s} \geq \exp\left(\min\{c_0, 1\} \frac{D}{2}\right).$$
Theorem

Let $\Pi(\cdot|Z^{(N)})$ arise from prior $N(0, \Sigma_{\alpha})$, $b = \alpha/d - 1/2 > 0$ and $D/N \approx \kappa > 0$. There exists $s_b \in (0, 1/2)$ s.t.:

i) The average log-likelihood is unimodal with mode 0, locally log-concave near 0, radially symmetric, Lipschitz continuous and decreasing in $\|\theta\|_{\mathbb{R}^D}$ on $\mathbb{R}^D$.

ii) For any $r > 0$ fixed and 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 $s = s_bN^{-b}$, we have $\Pi(B_s|Z^{(N)}) \xrightarrow{N \to \infty} 1$ in probability.
A general hitting time lower bound

**Theorem**

Let $\Pi(\cdot|Z^{(N)})$ arise from prior $N(0, \Sigma_\alpha)$, $b = \alpha/d - 1/2 > 0$ and $D/N \approx \kappa > 0$. There exists $s_b \in (0, 1/2)$ s.t.:

i) The average log-likelihood is unimodal with mode 0, locally log-concave near 0, radially symmetric, Lipschitz continuous and decreasing in $\|\theta\|_{\mathbb{R}^D}$ on $\mathbb{R}^D$.

ii) For any $r > 0$ fixed and 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 $s = s_bN^{-b}$, we have $\Pi(B_s|Z^{(N)}) \xrightarrow{N \to \infty} 1$ in probability.

iv) There exist $\varepsilon, C > 0$ s.t. for any Markov kernels $\mathcal{P}_N$ on $\mathbb{R}^D$ and associated chains $(\vartheta_k)$ satisfying Condition (A) for $\eta_N \in (0, s_bN^{-b})$, we can find an initialiser $\vartheta_0 \in \Theta_{N^{-b}, \varepsilon N^{-b}}$ s.t. w.h.p. the hitting time $\tau_{B_s}$ for $\vartheta_k$ to reach $B_s$ is

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

In fact we can take $\vartheta_0 \sim \mu|\Theta_{N^{-b}, \varepsilon N^{-b}}$ with $\mu = \Pi(\cdot|Z^{(N)})$. 
Let $\vartheta_k$ denote the MALA Markov chain with step size $\gamma$. Assume the setting of the general theorem with a $N(0, \Sigma_\alpha)$ prior. Then there exist some constant $c_1, c_2, \varepsilon > 0$ such that whenever

$$\gamma \leq c_1 N^{-1-b-2\alpha},$$

there is an initialisation point $\vartheta_0 \in \Theta_{N^{-b}, \varepsilon N^{-b}}$, such that with high probability under the data and the Markov chain,

$$\tau_{B_s} \geq \exp \left( c_2 D \right),$$

while still $\Pi(B_s | Z^{(N)}) \xrightarrow{N \to \infty} 1$ in probability.

So the MCMC outputs act effectively as a pure random number generator, not informed by the data likelihood.
A hardness result for pCN

**Hitting time lower bound II**

Let $\vartheta_k$ denote the pCN Markov chain with ‘step size’ $\beta$. For the $N(0, \mathcal{K})$-prior with $\mathcal{K} = \sum_\alpha$ for $\alpha > d/2$, let $\mathcal{G}$ be as in the previous Theorem.

Then there exist constants $c_1, c_2, \varepsilon > 0$ such that if $\beta \leq c_1 N^{-1 - 2b}$ there is an initialisation point $\vartheta_0 \in \Theta_{N^{-b}, \varepsilon N^{-b}}$ (or $\vartheta_0 \sim \mu|\Theta_{N^{-b}, \varepsilon N^{-b}}$) s.t. the hitting time

$$\tau_{B_s} = \inf\{k : \vartheta_k \in B_s\}$$

satisfies with high probability

$$\tau_{B_s} \geq \exp\left(c_2 D\right)$$

while still $\Pi(B_s|Z^{(N)}) \xrightarrow{N \to \infty} 1$ in probability.

This implies that the *dimension-independent* ‘spectral gaps’ from Hairer, Stuart & Vollmer (2014) exhibit exponential dependence on Lipschitz constants of $\Pi(\cdot|Z^{(N)})$. 
POLYNOMIAL TIME POSTERIOR COMPUTATION VIA GRADIENT STABILITY AND LOG-CONCAVE APPROXIMATION

[Joint work with S. Wang (2022), and also J. Bohr (2023)]
The linearisation of $G$

For bounded perturbations, let $G'_\theta : \Theta \to L^2$ be the linear operator s.t.

$$
\|G(\theta + h) - G(\theta) - G'_\theta[h]\|_{L^2} = o(\|h\|) \to 0.
$$
The linearisation of $G$

For bounded perturbations, let $G_\theta' : \Theta \to L^2$ be the linear operator s.t.

$$\|G(\theta + h) - G(\theta) - G_\theta'[h]\|_{L^2} = o(\|h\|) \to 0.$$ 

We require a stability inequality quantifying the ‘local injectivity’ of $G_\theta'$ at $\theta_0$.

‘Gradient stability’

Assume at $\theta_0 \in \mathbb{R}^D$ that for some $\kappa_0 \geq 0$,

$$\|G_{\theta_0}'[h]\|_{L^2}^2 \geq D^{-\kappa_0} \|h\|^2 \quad \forall h \in \mathbb{R}^D.$$
The linearisation of $G$

For bounded perturbations, let $G'_\theta : \Theta \to L^2$ be the linear operator s.t.

$$\|G(\theta + h) - G(\theta) - G'_\theta[h]\|_{L^2} = o(\|h\|) \to 0.$$  

We require a stability inequality quantifying the ‘local injectivity’ of $G'_\theta$ at $\theta_0$.

‘Gradient stability’

Assume at $\theta_0 \in \mathbb{R}^D$ that for some $\kappa_0 \geq 0$,

$$\|G'_{\theta_0}[h]\|_{L^2}^2 \gtrsim D^{-\kappa_0} \|h\|^2 \quad \forall h \in \mathbb{R}^D.$$  

Here $\kappa_0 > 0$ depends on the ‘local ill-posedness’ of $G$.

- For the Schrödinger equation: $\kappa_0 = 4/d$
- For Darcy’s problem $\kappa = 6/d$
- For (non-Abelian) X-ray transforms $\kappa = 1/2$
Local ‘average curvature’ in nonlinear models

The lack of log-concavity of the posterior manifests itself in \( (\ell = \ell_1) \)

\[
-\nabla^2 \ell(\theta, Z) = [\nabla G(\theta)(X)][\nabla G(\theta)(X)]^T + [G(\theta)(X) - Y] \nabla^2 [G(\theta)(X)].
\]

For \( Y, X \) fixed there is no reason why \(-\nabla^2 \ell \) should be (even only locally) convex.
Local ‘average curvature’ in nonlinear models

The lack of log-concavity of the posterior manifests itself in \((\ell = \ell_1)\)

\[-\nabla^2 \ell(\theta, Z) = [\nabla \mathcal{G}(\theta)(X)][\nabla \mathcal{G}(\theta)(X)]^T + [\mathcal{G}(\theta)(X) - Y]\nabla^2[\mathcal{G}(\theta)(X)].\]

For \(Y, X\) fixed there is no reason why \(-\nabla^2 \ell\) should be (even only locally) convex.

However the ‘average’ Hessian computed under the sampling distribution \(P^N_{\theta_0}\)

satisfies near \(\theta_0\) and for \(||h||_{\mathbb{R}^D} \leq 1\) (and appropriate norm \(||\cdot||_*\) )

\[h^T E_{\theta_0}[−\nabla^2 \ell(\theta, Z)]h = ||h^T \nabla \mathcal{G}(\theta)||_{L^2}^2 + O(||\mathcal{G}(\theta) - \mathcal{G}(\theta_0)||_*).\]

‘Gradient stability’ controls the first term since \(h^T \nabla \mathcal{G}(\theta) = \mathcal{G}'_0[h], \ h \in \mathbb{R}^D.\)
Local ‘average curvature’ in nonlinear models

The lack of log-concavity of the posterior manifests itself in \((\ell = \ell_1)\)

\[-\nabla^2 \ell(\theta, Z) = [\nabla G(\theta)(X)][\nabla G(\theta)(X)]^T + [G(\theta)(X) - Y] \nabla^2 [G(\theta)(X)].\]

For \(Y, X\) fixed there is no reason why \(-\nabla^2 \ell\) should be (even only locally) convex.

However the ‘average’ Hessian computed under the sampling distribution \(P_{\theta_0}^N\) satisfies near \(\theta_0\) and for \(\|h\|_{\mathbb{R}^D} \leq 1\) (and appropriate norm \(\|\cdot\|_*\))

\[h^T E_{\theta_0}[-\nabla^2 \ell(\theta, Z)] h = \|h^T \nabla G(\theta)\|_{L_2}^2 + O(\|G(\theta) - G(\theta_0)\|_*).\]

‘Gradient stability’ controls the first term since \(h^T \nabla G(\theta) = G'_0[h], \ h \in \mathbb{R}^D.\)

Hypothesis (local average convexity of \(-\ell_N/N\))

\[\inf_{\theta \in B} \lambda_{\min}(E_{\theta_0}[-\nabla^2 \ell(\theta, Z)]) \geq c_{\min} > 0\]

on some neighbourhood \(B\) of \(\theta_0\), whose size needs to be quantified.

For the PDE examples, \(G\) is sufficiently smooth that gradient stability implies the last condition for appropriate neighbourhoods \(B\) of radius \(D^{-w}, w > 0.\)
Concentration of measure: Local average curvature extends to the observed likelihood function $\ell_N$ (empirical measures concentrate in high dimensions, Talagrand (2014), Giné & N (2016), Vershynin (2018)).

**Theorem**

With high $P_{\theta_0}^N$-probability and for $D \lesssim N^b$ some $b > 0$, one has,

$$\inf_{\theta \in B} \lambda_{\min}\left[-\nabla^2 \ell_N(\theta, Z)\right] \geq Nc_{\min} > 0.$$
Concentration of measure: Local average curvature extends to the observed likelihood function $\ell_N$ (empirical measures concentrate in high dimensions, Talagrand (2014), Giné & N (2016), Vershynin (2018)).

**Theorem**

With high $P_{\theta_0}^N$-probability and for $D \lesssim N^b$ some $b > 0$, one has,

$$\inf_{\theta \in B} \lambda_{\min}[-\nabla^2 \ell_N(\theta, Z)] \geq Nc_{\min} > 0.$$  

Under global injectivity hypotheses for $G$ the posterior is statistically consistent (cf. Nickl (2023)) and puts its mass precisely in the region $B$ of log-concavity near $\theta_0$. We then ‘concavify’ $\mathbb{P}(\cdot|Z^{(N)})$ near $\theta_0$ by a proxy measure $\tilde{\mathbb{P}}(\cdot|Z^{(N)})$.
Roadmap to exploiting local average convexity

**Concentration of measure:** Local average curvature extends to the observed likelihood function $\ell_N$ (empirical measures concentrate in high dimensions, Talagrand (2014), Giné & N (2016), Vershynin (2018)).

**Theorem**

With high $P^N_{\theta_0}$-probability and for $D \lesssim N^b$ some $b > 0$, one has,

$$\inf_{\theta \in \mathcal{B}} \lambda_{\min} \left[ -\nabla^2 \ell_N(\theta, Z) \right] \geq Nc_{\min} > 0.$$

Under global injectivity hypotheses for $G$ the posterior is statistically consistent (cf. Nickl (2023)) and puts its mass precisely in the region $\mathcal{B}$ of log-concavity near $\theta_0$. We then ‘concavify’ $\Pi(\cdot|Z^{(N)})$ near $\theta_0$ by a proxy measure $\tilde{\Pi}(\cdot|Z^{(N)})$.

**Theorem**

Assuming local and global regularity of $G$ we have whp under the data that the proxy measure $\tilde{\Pi}(\cdot|Z^{(N)})$ is strongly globally-log-concave and satisfies

$$W_2^2(\tilde{\Pi}(\cdot|Z^{N}), \Pi(\cdot|Z^{N})) \leq \exp(-N\bar{b}), \quad \bar{b} > 0.$$
Consider computation of the high-dimensional Bochner integral

$$E_{\Sigma}^{\Pi}[\theta|Z^{(N)}] = \int_{\mathbb{R}^D} \theta d\Pi(\theta|Z^{(N)})$$

under appropriate assumptions on $D, G, \Pi, \theta_0$, covering our PDE examples.
Polynomial-time algorithms for posterior mean vectors

Consider computation of the high-dimensional Bochner integral

\[ E^\Pi[\theta|Z^{(N)}] = \int_{\mathbb{R}^D} \theta d\Pi(\theta|Z^{(N)}) \]

under appropriate assumptions on \( D, \mathcal{G}, \Pi, \theta_0 \), covering our PDE examples. We assume an initialiser into the region where average curvature holds, and then run ULA on the proxy measure \( \tilde{\Pi}(\cdot|Z^{(N)}) \).
Consider computation of the high-dimensional Bochner integral

$$E^\Pi[\theta|Z^{(N)}] = \int_{\mathbb{R}^D} \theta d\Pi(\theta|Z^{(N)})$$

under appropriate assumptions on $D, \mathcal{G}, \Pi, \theta_0$, covering our PDE examples. We assume an initialiser into the region where average curvature holds, and then run ULA on the proxy measure $\tilde{\Pi}(\cdot|Z^{(N)})$.

**Theorem**

*For any precision level $\varepsilon \geq N^{-P}$, there exists a (‘warm start’) sampling algorithm with polynomial computational cost*

$$O(N^{b_1} D^{b_2} \varepsilon^{-b_3}) \quad (b_1, b_2, b_3 > 0),$$
Polynomial-time algorithms for posterior mean vectors

Consider computation of the high-dimensional Bochner integral

$$E^\Pi[\theta|Z^{(N)}] = \int_{\mathbb{R}^D} \theta d\Pi(\theta|Z^{(N)})$$

under appropriate assumptions on $D, G, \Pi, \theta_0$, covering our PDE examples.
We assume an initialiser into the region where average curvature holds, and then run ULA on the proxy measure $\tilde{\Pi}(\cdot|Z^{(N)})$.

**Theorem**

*For any precision level $\varepsilon \geq N^{-P}$, there exists a (‘warm start’) sampling algorithm with polynomial computational cost

$$O(N^{b_1} D^{b_2} \varepsilon^{-b_3}) \quad (b_1, b_2, b_3 > 0),$$

and whose output $\hat{\theta}_\varepsilon$ satisfies that with high probability

$$\|\hat{\theta}_\varepsilon - E^\Pi[\theta|Z^{(N)}]\|_{\mathbb{R}^D} \leq \varepsilon \quad \text{as well as} \quad \|\hat{\theta}_\varepsilon - \theta_0\|_{\mathbb{R}^D} \leq \varepsilon$$*

