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

Statistical inverse regression models

Consider statistical observations arising as random vectors

$$Y_i = \mathscr{G}_\theta(X_i) + \varepsilon_i, \quad \varepsilon_i \sim^{i.i.d.} \mathcal{N}_V(0, I),$$

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

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

Statistical inverse regression models

Consider statistical observations arising as random vectors

$$Y_i = \mathscr{G}_\theta(X_i) + \varepsilon_i, \quad \varepsilon_i \sim^{i.i.d.} \mathcal{N}_V(0, I),$$

- The X_i are covariates drawn iid from some law λ on a d-dimensional domain $\mathcal{X}$.
- The Y_i are 'response' variables in a vector space V of finite $dim(V)$, 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

$$\{\mathscr{G}_\theta : \theta \in \Theta\}, \quad \mathscr{G}_\theta : \mathcal{X} \to V,$$

are indexed by the high-dimensional parameter

$$\theta \in \Theta = \mathbb{R}^D$$

arising from the discretisation of a function space in some basis. **We take asymptotics $D, N \to \infty$, possibly $D/N \to \kappa > 0$.**

PDE model examples – Non-linear inverse problems

• The dependence $\theta \mapsto \mathscr{G}(\theta)$ is non-linear and we have in mind coefficient to solution maps of partial differential equations (PDEs).

PDE model examples – Non-linear inverse problems

• The dependence $\theta \mapsto \mathscr{G}(\theta)$ 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 $u = u_\theta = \mathscr{G}_\theta$ of the elliptic PDE

$$\begin{aligned} \nabla \cdot (f_\theta \nabla u) &= g \text{ in } \mathcal{X}, \\ u &= 0 \text{ on } \partial\mathcal{X}. \end{aligned}$$

PDE model examples – Non-linear inverse problems

• The dependence $\theta \mapsto \mathscr{G}(\theta)$ 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 $u = u_\theta = \mathscr{G}_\theta$ of the elliptic PDE

$$\begin{aligned} \nabla \cdot (f_\theta \nabla u) &= g \text{ in } \mathcal{X}, \\ u &= 0 \text{ on } \partial\mathcal{X}. \end{aligned}$$

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

PDE model examples – Non-linear inverse problems

• The dependence $\theta \mapsto \mathscr{G}(\theta)$ 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 $u = u_\theta = \mathscr{G}_\theta$ of the elliptic PDE

$$\begin{aligned} \nabla \cdot (f_\theta \nabla u) &= g \text{ in } \mathcal{X}, \\ u &= 0 \text{ on } \partial\mathcal{X}. \end{aligned}$$

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

• The 'forward' map $\theta \to \mathscr{G}_\theta$ can be 'evaluated' by numerical PDE methods (finite elements, etc.).

Penalised Least Squares / Tikhonov Regularisation

- A natural approach (Gauß, Tikhonov) is to minimise a least squares fit

$$Q_N(\theta) = \sum_{i=1}^{N} |Y_i - \mathscr{G}_\theta(X_i)|_V^2 + \lambda \cdot pen(\theta), \quad \lambda > 0,$$

over $\theta \in \mathbb{R}^D$. The penalty term is also called a 'regulariser'.

Penalised Least Squares / Tikhonov Regularisation

- A natural approach (Gauß, Tikhonov) is to minimise a least squares fit

$$Q_N(\theta) = \sum_{i=1}^{N} |Y_i - \mathscr{G}_\theta(X_i)|_V^2 + \lambda \cdot pen(\theta), \quad \lambda > 0,$$

  over $\theta \in \mathbb{R}^D$. The penalty term is also called a 'regulariser'.

- Classical choices are Sobolev type norms

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

  or related ℓ_1-type penalties/TV-type norms.

Penalised Least Squares / Tikhonov Regularisation

- A natural approach (Gauß, Tikhonov) is to minimise a least squares fit

$$Q_N(\theta) = \sum_{i=1}^{N} |Y_i - \mathscr{G}_\theta(X_i)|_V^2 + \lambda \cdot pen(\theta), \quad \lambda > 0,$$

over $\theta \in \mathbb{R}^D$. The penalty term is also called a 'regulariser'.

- Classical choices are Sobolev type norms

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

or related ℓ_1-type penalties/TV-type norms.

- As $\mathscr{G}$ is non-linear in θ, the map Q_N is not convex on $\mathbb{R}^D$.

Penalised Least Squares / Tikhonov Regularisation

- A natural approach (Gauß, Tikhonov) is to minimise a least squares fit

$$Q_N(\theta) = \sum_{i=1}^{N} |Y_i - \mathscr{G}_\theta(X_i)|_V^2 + \lambda \cdot pen(\theta), \quad \lambda > 0,$$

  over $\theta \in \mathbb{R}^D$. The penalty term is also called a 'regulariser'.

- Classical choices are Sobolev type norms

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

  or related ℓ_1-type penalties/TV-type norms.

- As $\mathscr{G}$ is non-linear in θ, the map Q_N is not convex on $\mathbb{R}^D$.

- The algorithmic runtime $= \{\#$ required evaluations of $\mathscr{G}(\theta)\}$ to compute such optimisers may scale exponentially in dimension D and sample size N.

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

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

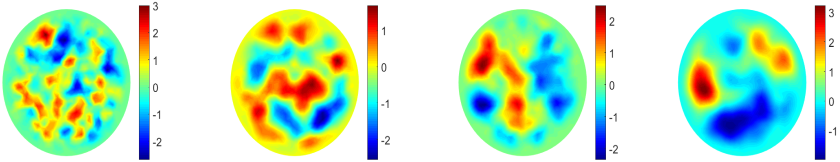
modelling α-regular stationary random fields θ over $\mathcal{Z}$, with RKHS $\mathcal{H} = H^\alpha$.

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

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

modelling α-regular stationary random fields θ 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 θ and then also $\mathscr{G}_\theta$ in inverse problems.

Bayesian Inversion with Gaussian priors

The 'log-likelihood' of our statistical regression model is

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

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

Bayesian Inversion with Gaussian priors

The 'log-likelihood' of our statistical regression model is

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

whose evaluation requires only computation of $\mathscr{G}(\theta)$ (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

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

whose evaluation requires only computation of $\mathscr{G}(\theta)$ (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

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

If a Markov chain (ϑ_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.

Random walk with Metropolis-Hastings adjustment (pCN)

Following Cotter, Roberts, Stuart & White (2013), Hairer, Stuart, Vollmer (2014):

pre-conditioned Crank-Nicolson (pCN) algorithm

Let $\Pi \sim N(0, \mathcal{K})$ on $\Theta = \mathbb{R}^D$. Fix $\delta > 0$ and initialise ϑ_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 $\Pi(\cdot|Z^{(N)})$.

Computation: gradient based MCMC

Discretised Langevin type algorithm (ULA)

Choose step size $\delta > 0$ and an initialiser ϑ_0. For $\xi_k \sim^{iid} \mathcal{N}(0, I)$ in $\mathbb{R}^D$, do:

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

Computation: gradient based MCMC

Discretised Langevin type algorithm (ULA)

Choose step size $\delta > 0$ and an initialiser ϑ_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.

Algorithmic complexity

Is posterior computation by MCMC possible in such non-linear regression problems **in polynomial run time** in dimension D and sample size (informativeness) N?

Algorithmic complexity

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

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

Algorithmic complexity

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).

Algorithmic complexity

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).

Sampling from log-concave targets

Bakry & Emery (1985), Dalayan (2017), Durmus & Moulines (2019)..

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** $\mathscr{G}$ and Gaussian priors as the posterior is then log-concave.

Sampling from log-concave targets

Bakry & Emery (1985), Dalayan (2017), Durmus & Moulines (2019)..

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** $\mathscr{G}$ and Gaussian priors as the posterior is then log-concave.

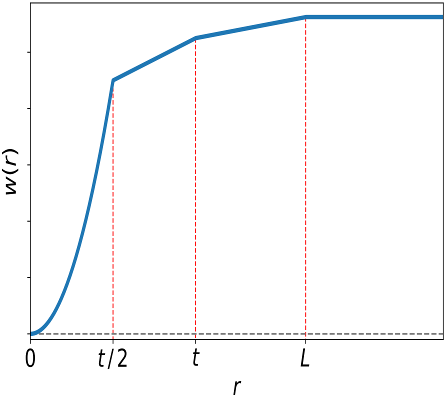
The theory extends to target measures satisfying a **log-Sobolev inequality** (e.g., Vempala & Wibisono (2021)), but the LSI-constants scale *exponentially* in bounded ('Holley-Stroock') perturbations.

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 α-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



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.

Setup and notation: $B_s, \Theta_{r,r+\varepsilon}$

Recall that for Whittle-Matern $N(0, \Sigma_\alpha)$-prior, the posterior on $\Theta = \mathbb{R}^D$ is

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

Setup and notation: $B_s, \Theta_{r,r+\varepsilon}$

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 $\theta_0 \in \mathbb{R}^D$,

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

and D-dimensional annuli

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

We consider non-intersecting 'inner' and 'outer' annuli $\Theta_{s,\eta}$ and $\Theta_{r,\varepsilon}$, with $s < \sigma$.

Setup and notation: $B_s, \Theta_{r,r+\varepsilon}$

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

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

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.

Jerrum's bottleneck

General hitting time bound

Consider any Markov chain $(\vartheta_k : k \in \mathbb{N})$ with invariant 'target' measure μ (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 ϑ_0 is started in the 'outer annulus' $\Theta_{\sigma,\varepsilon}$, drawn from the **conditional** distribution $\mu(\cdot|\Theta_{\sigma,\varepsilon})$, and denote by

$$\tau_s = \inf\{k : \vartheta_k \in \Theta_{s,\eta}\}$$

the **hitting time** of the Markov chain onto the intermediate annulus $\Theta_{s,\eta}$. Then

$$\Pr(\tau_s \leq K) \leq K e^{-\nu N}, \text{ for all } K > 0.$$

Large Deviation Landscape (Franz-Parisi functional)

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

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

Large Deviation Landscape (Franz-Parisi functional)

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

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

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 $\Theta_{1,1+\varepsilon}$.

Large Deviation Landscape (Franz-Parisi functional)

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

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

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 $\Theta_{1,1+\varepsilon}$.

For α-regular Gaussian process priors with $\eta = o(N^{-b}), b = b_{\alpha,d} > 0$, and $D/N \simeq \kappa > 0$, 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 $\sigma = 1$,

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

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 $\Theta_{1,1+\varepsilon}$.

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

$$\frac{1}{N}\log\frac{\Pi(\Theta_{s,\eta})}{\Pi(\Theta_{1,\varepsilon})} \leq -\nu, \quad \nu > 0, \quad \text{some } \varepsilon > 0.$$

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.**

Genius in a bottle and step sizes of Markov chains

Let $\mathcal{P}_N(\theta, A)$ 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 (ϑ_k).

Condition (A)

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

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

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

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.

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.:

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 $\mathbb{R}^D$.

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 $\{\theta : \|\theta\|_{\mathbb{R}^D} \geq rN^{-b}\}$.

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 $\Pi(B_s|Z^{(N)}) \xrightarrow{N\to\infty} 1$ 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 $D/N \simeq \kappa > 0$. 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 $\Pi(B_s|Z^{(N)}) \xrightarrow{N\to\infty} 1$ in probability.

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

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

In fact we can take $\vartheta_0 \sim \mu_{|\Theta_{N^{-b}, \varepsilon N^{-b}}}$ with $\mu = \Pi(\cdot|Z^{(N)})$.

A hardness result for MALA

Hitting time lower bound I

Let ϑ_k denote the MALA Markov chain with step size γ. 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 ϑ_k denote the pCN Markov chain with 'step size' β. For the $N(0, \mathcal{K})$-prior with $\mathcal{K} = \Sigma_\alpha$ 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 o 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 $\mathscr{G}$

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

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

The linearisation of $\mathscr{G}$

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

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

We require a stability inequality quantifying the 'local injectivity' of $\mathscr{G}'_\theta$ at θ_0.

'Gradient stability'

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

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

The linearisation of $\mathscr{G}$

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

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

We require a stability inequality quantifying the 'local injectivity' of $\mathscr{G}'_\theta$ at θ_0.

'Gradient stability'

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

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

Here $\kappa_0 > 0$ depends on the 'local ill-posedness' of $\mathscr{G}$.

- 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 $(\ell = \ell_1)$

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

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 ($\ell = \ell_1$)

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

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 θ_0 and for $\|h\|_{\mathbb{R}^D} \leq 1$ (and appropriate norm $\|\cdot\|_*$)

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

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

Local 'average curvature' in nonlinear models

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

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

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 θ_0 and for $\|h\|_{\mathbb{R}^D} \leq 1$ (and appropriate norm $\|\cdot\|_*$)

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

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

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

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

on some neighbourhood $\mathcal{B}$ of θ_0, whose size *needs to be quantified*.

For the PDE examples, $\mathscr{G}$ is sufficiently smooth that gradient stability implies the last condition for appropriate neighbourhoods $\mathcal{B}$ of radius $D^{-w}, w > 0$.

Roadmap to exploiting local average convexity

Concentration of measure: Local average curvature *extends to the observed likelihood* function ℓ_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,

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

Roadmap to exploiting local average convexity

Concentration of measure: Local average curvature *extends to the observed likelihood* function ℓ_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,

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

Under global injectivity hypotheses for $\mathscr{G}$ the posterior is statistically consistent (cf. Nickl (2023)) and puts its mass precisely in the region $\mathcal{B}$ of log-concavity near θ_0. We then 'concavify' $\Pi(\cdot|Z^{(N)})$ near θ_0 by a proxy measure $\tilde{\Pi}(\cdot|Z^{(N)})$.

Roadmap to exploiting local average convexity

Concentration of measure: Local average curvature *extends to the observed likelihood* function ℓ_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,

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

Under global injectivity hypotheses for $\mathscr{G}$ the posterior is statistically consistent (cf. Nickl (2023)) and puts its mass precisely in the region $\mathcal{B}$ of log-concavity near θ_0. We then 'concavify' $\Pi(\cdot|Z^{(N)})$ near θ_0 by a proxy measure $\tilde{\Pi}(\cdot|Z^{(N)})$.

Theorem

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

$$W_2^2\big(\tilde{\Pi}(\cdot|Z^N), \Pi(\cdot|Z^N)\big) \leq \exp(-N^{\bar{b}}), \;\; \bar{b} > 0.$$

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 $D, \mathscr{G}, \Pi, \theta_0$, 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 $D, \mathscr{G}, \Pi, \theta_0$, covering our PDE examples.

We assume an initialiser into the region where average curvature holds, and then run ULA on the proxy measure $\tilde{\Pi}(\cdot|Z^{(N)})$.

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 $D, \mathscr{G}, \Pi, \theta_0$, 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 $D, \mathscr{G}, \Pi, \theta_0$, 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

$$\left\|\hat{\theta}_\varepsilon - E^{\Pi}[\theta|Z^{(N)}]\right\|_{\mathbb{R}^D} \leq \varepsilon \;\; \textit{as well as} \; \left\|\hat{\theta}_\varepsilon - \theta_0\right\|_{\mathbb{R}^D} \leq \varepsilon$$

References

R. Nickl, *Bayesian non-linear statistical inverse problems*, Zürich Lectures in Advanced Mathematics (EMS press), (2023)

R. Nickl, S. Wang, On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms, *J. Eur. Math. Soc.*, (2022)

A. Bandeira, A. Maillard, R. Nickl, S. Wang, On free energy barriers in Gaussian priors and failure of cold start MCMC for high-dimensional unimodal distributions, *Phil. Trans. Roy. Soc. A*, (2023)

J. Bohr, R. Nickl, On log-concave approximations of high-dimensional posterior measures and stability properties in non-linear inverse problems. *Ann. Inst. H. Poincaré (Probab. Statist.)*, (2023).