On polynomial-time mixing for high-dimensional MCMC in non-linear regression

Sven Wang (HU Berlin)

Research unit seminar Mathematical Statistics in the Information Age"

November 3, 2023

Joint work with:
Richard Nickl (Cambridge), A.S. Bandeira (ETH), Antoine Maillard (ETH)

arxiv:2009.05298 (J. Eur. Math. Soc., 2022)
arxiv:2209.02001 (Phil. Trans. R. Soc. A, 2023)

Bayesian computation

In this talk, we consider sampling from a D-dimensional **posterior distribution**

$$\pi(\theta|Z^{(N)}) \propto e^{\ell_N(\theta)}\pi(\theta), \quad \theta \in \mathbb{R}^D,$$

using **Markov chain Monte Carlo (MCMC)**.

- π is prior distribution on $\mathbb{R}^D$.
- ℓ_N is log-likelihood function at **sample size** N.
- D is model dimension.

Sampling tasks are ubiquitous in **high-dimensional Bayesian statistics, inverse problems, data assimilation, etc.**

This talk: Synthesis of statistical and computational theory.

Main computational question

In complex models,

- D may grow with N ('high-dimensional').
- $-\ell_N$ may be complex (e.g. non-convex, multi-modal).

> **Fundamental question:** Are high-dimensional Bayesian posterior distributions feasible to compute and if so, when?

Goals:

(A) Generate random variables $\vartheta \in \mathbb{R}^D$ with law $\mathcal{L}(\vartheta) \approx \Pi(\theta|Z^{(N)})$,
(B) Compute 'aspects' of the posterior

- Functionals: $\int_\Theta H(\theta) d\Pi(\theta|Z^{(N)}) \in \mathbb{R}$.
- Posterior mean: $\int_\Theta \theta d\Pi(\theta|Z^{(N)}) \in \mathbb{R}^D$,
- MAP estimate: $\hat{\theta}_{MAP} \in \arg\max_\theta \pi(\theta|Z^{(N)})$?

In this talk: 'feasible' computational cost = polynomial time.

Popular approach: Markov Chain Monte Carlo

Basic idea: Generate ergodic Markov chain $(\vartheta_k : k = 0, 1, ...)$ on $\mathbb{R}^D$ such that

$\Pi(\cdot|Z^{(N)})$ is the unique invariant distribution for (ϑ_k).

Tasks **(A)** and **(B)** are then naturally addressed:

(A) To approximate the posterior distribution, after some *mixing time* J_{mix},

$$\mathcal{L}(\vartheta_k) \approx \Pi(\cdot|Z^{(N)}) \quad \text{for any} \quad k \geq J_{mix}.$$

(B) To compute integrals, we can take *ergodic averages*,

$$\frac{1}{J} \sum_{k=J_{mix}+1}^{J_{mix}+J} H(\vartheta_k) \approx \int_\Theta H(\theta) d\Pi(\theta|Z^{(N)}), \quad J \in \mathbb{N}.$$

Question rephrased: How do J_{mix}, J depend on D and N?

Inverse (regression) problem

Let $\mathcal{O} \subseteq \mathbb{R}^d$ (bounded, $d \leq 3$) and consider a forward map

$$\mathcal{G} : \mathbb{R}^D \to C(\mathcal{O}), \quad D \in \mathbb{N}.$$

We consider (non-linear) $\mathcal{G}$ arising from some PDE model.

Regression data: We observe $Z^{(N)} = (Y_i, X_i)_{i=1}^N$, where

$$Y_i = \mathcal{G}(\theta)(X_i) + \varepsilon_i, \quad i = 1, ..., N,$$

and $\varepsilon_i \sim^{i.i.d.} N(0,1)$, $X_i \sim^{i.i.d.} Uniform(\mathcal{O})$.

Prominent examples:

- Inverse problems [Stuart (2010)]
- Data assimilation [Majda & Harlim (2012), Reich & Cotter (2015)]

Bayesian inference with Gaussian priors

Gaussian prior: Suppose $\theta \sim N(0, \Sigma)$.

Posterior density:

$$\pi(\theta|Z^{(N)}) \propto \exp\Big[-\frac{1}{2}\sum_{i=1}^{N}\big(Y_i - \mathcal{G}(\theta)(X_i)\big)^2 - \frac{1}{2}\theta^T\Sigma^{-1}\theta\Big].$$

Characteristics:

- **Non-log-concavity.** For *nonlinear* $\mathcal{G}$, $-\ell_N$ may be *non-convex*.
- **Spiked-ness.** $\pi(\cdot|Z^{(N)})$ may be multimodal with local optima of depth $O(N)$.
- **Posterior consistency:** the data becomes informative as $N \to \infty$ (and thus $D \to \infty$).

Curse of dimensionality? Sampling may require exponential (in D, N) computation [Bickel et al (2008), Rebeschini & van Handel (2015), Yang, Wainwright & Jordan (2016)]

Theoretical guarantees for large D

I. Preconditioned Crank-Nicolson

- [Hairer, Stuart and Vollmer (2014)] prove *dimension-independent* 'spectral gap', but exponential dependence on other parameters.

II. Langevin Monte Carlo

- Log-concave: [Dalalyan (2017), Durmus & Moulines (2017,2019)] derive bounds with *polynomial dependence* on relevant parameters.
- Non-log-concave: [Vempala & Wibisono (2019)] use log-Sobolev assumptions, [Ma et al. (2018) use convexity outside some Euclidean ball.

How to break the **exponential vs. polynomial barrier**? In the statistical setting, have **additional structure**:

- Bayesian posterior contraction [Ghosal & van der Vaart '17]. How much can computation be localized?
- Bernstein-von Mises phenomenon / asymptotic normality, even for non-linear problems [e.g. Nickl '17].

I. Upper bounds

A concrete example: Steady-state Schrödinger equation

Let $\mathcal{O} \subseteq \mathbb{R}^d$ be bounded ($d \leq 3$) and $g : \partial\mathcal{O} \to (g_{min}, \infty)$, $g_{min} > 0$, be *given*.

PDE solution map

$$G : f \mapsto u_f \equiv u, \quad \text{with} \begin{cases} \frac{\Delta}{2} u - fu = 0 & \text{on } \mathcal{O}, \\ u = g & \text{on } \partial O. \end{cases}$$

- For sufficiently regular $f : \mathcal{O} \to [0, \infty)$, a unique solution $u_f \in C^2(\mathcal{O})$ exists by elliptic PDE theory.
- G is nonlinear, as seen from the Feynman-Kac formula

$$u_f(x) = E_x\left[g\left(X_{\tau_{\mathcal{O}}}\right) e^{-\int_0^{\tau_{\mathcal{O}}} f(X_s) ds}\right],$$

  where $(X_t : t \geq 0)$ is a d-dimensional Brownian motion started at $x \in \mathcal{O}$ with exit time $\tau_{\mathcal{O}}$.
- Appears e.g. in photoacoustic tomography (PAT) [Bal and Uhlmann (2009)].

The forward map $\mathcal{G}$

We fix a parameterisation

$$\theta \mapsto f_\theta, \qquad f_\theta \equiv \Phi \circ \Big(\sum_{k=1}^{D} \theta_k e_k \Big).$$

- Here $(e_k : k \geq 1)$ are *D eigenfunctions* of the Dirichlet Laplacian.
- $\Phi : \mathbb{R} \to (0, \infty)$ is a 'regular' link function *ensuring positivity*.

The forward map is then given by

$$\mathcal{G} : \mathbb{R}^D \to C^2(\mathcal{O}), \quad \theta \mapsto G(f_\theta) \equiv u_{f_\theta}.$$

In what follows, we assume $Z^{(N)}$ is drawn from $P^N_{\theta_0}$ for some ground truth θ_0 (not necessarily $\theta_0 \in \mathbb{R}^D$).

Existence of polynomial-time algorithm

Our first main result regards computation of the *posterior mean*,

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

Theorem (Nickl and W, 2020)
Grant mild regularity assumptions on model dimension D, the prior Π and the ground truth θ_0. For any $P > 0$ and precision level $\varepsilon \geq N^{-P}$, there exists a 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 (under the joint law of the data $Z^{(N)}$ and the randomness of the algorithm),

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

Description of algorithm

The algorithm is a Langevin-type Markov chain $(\vartheta_k : k \geq 0)$.

$$\begin{cases} \vartheta_0 & = \theta_{init}, \\ \vartheta_{k+1} & = \vartheta_k + \gamma \nabla \log \tilde{\pi}(\vartheta_k | Z^{(N)}) + \sqrt{2\gamma}\xi_{k+1}, \end{cases}$$

- The **initialiser** θ_{init} is computable in polynomial time,
- $(\xi_k : k \geq 1)$ are $N(0, I_{D \times D})$ random variables,
- $\tilde{\pi}(\cdot | Z^{(N)})$ is a **log-concave 'proxy' posterior distribution**,

$$\tilde{\pi}(\theta | Z^{(N)}) \propto e^{\tilde{\ell}_N(\theta)} \pi(\theta),$$

  where $\tilde{\ell}_N$ is a surrogate likelihood $\tilde{\ell}_N = \ell_N$ for $\|\theta - \theta_{0,D}\|_{\mathbb{R}^D} \lesssim D^{-4/d}$.
- $\tilde{\pi}(\theta | Z^{(N)})$ is constructed from the data in poly-time.

Hypotheses for the next theorems

- **Precision level:** Let $\varepsilon \geq N^{-P}$ (any $P > 0$ fixed).
- **Dimension constraint:**

$$D \lesssim N^{d/(2\alpha+d)}, \quad \alpha > 6.$$

- **Bias:** Suppose that θ_0 is sufficiently well-approximated by $\theta_{0,D} \in \mathbb{R}^D$.
- **Step size:** For some $a > 0$ (e.g. when $d = 3$, $a \approx 8$), set

$$\gamma = \gamma_{\varepsilon,D,N} \simeq \varepsilon^2 D^{-a} N^{-1}.$$

Concentration inequalities for ergodic averages

For a 1-Lipschitz function, $H : \mathbb{R}^D \to \mathbb{R}$, let

$$\hat{\mu}(H) = \frac{1}{J} \sum_{k=J_{mix}+1}^{J_{mix}+J} H(\vartheta_k).$$

Theorem (Nickl and W, 2020)
There exist constants $g_{D,N,\varepsilon} = O(D^{b_1} N^{b_2} \varepsilon^{-b_3})$, $b_1, b_2, b_3 > 0$ *such that for* $J_{mix} \geq g_{D,N,\varepsilon}$ *and with high probability under the data,*

$$\mathbf{P}_{MCMC}\left(\left|\hat{\mu}(H) - E^{\Pi}[H|Z^{(N)}]\right| \geq \varepsilon\right) \lesssim \exp(-J/g_{D,N,\varepsilon}),$$

Hence, there exists $\eta > 0$ such that with high $P_{\theta_0}^N \times \mathbf{P}_{MCMC}$-probability and polynomially many iterates,

$$\left\|\bar{\theta}_{J_{mix}}^J - \theta_0\right\|_{\mathbb{R}^D} \lesssim N^{-\eta}.$$

Convergence in Wasserstein distance

For probability measures μ_1, μ_2 on $\mathbb{R}^D$, define the Wasserstein-2 distance

$$W_2^2(\mu_1,\mu_2) := \inf_{\nu\in\Gamma(\mu_1,\mu_2)} \int_{\mathbb{R}^D} \|\theta_1-\theta_2\|_{\mathbb{R}^D}^2 \, d\nu(\theta_1,\theta_2).$$

Theorem (Nickl and W, 2020)
With high probability, the Markov chain (ϑ_k) *with step size* $\gamma_\varepsilon > 0$ *satisfies that for all* $k \geq 1$,

$$W_2^2\big(\mathcal{L}(\vartheta_k), \Pi(\cdot|Z^{(N)})\big) \lesssim D^{2\alpha/d}\big(1 - c\gamma_\varepsilon N D^{-4/d}\big)_+^k + \varepsilon^2,$$

and $W_2^2\big(\mathcal{L}(\vartheta_k), \Pi(\cdot|Z^{(N)})\big) \leq 2\varepsilon^2$ *for* $k \geq k_{mix} = O(D^{b_1} N^{b_2} \varepsilon^{-b_3})$.

- The term $c\gamma_\varepsilon N D^{-4/d}$ can be thought of as a 'spectral gap'.
- The error incurred by *Euler discretisation* and *proxy construction* is $\leq \varepsilon^2$.

Theorem (Nickl and W, 2020)
Consider the gradient descent algorithm

$$\vartheta_0 = \theta_{init}, \quad \vartheta_{k+1} = \vartheta_k + \gamma_\varepsilon \nabla \log \tilde{\pi}(\vartheta_k | Z^{(N)}).$$

Then, with high $P^N_{\theta_0}$-probability, we have

$$\left\| \vartheta_k - \hat{\theta}_{MAP} \right\|_{\mathbb{R}^D} \lesssim \left(1 - \frac{c}{D^{4/d} \gamma_\varepsilon} \right)^k \quad \text{for all } k \geq 1.$$

Moreover, for some constant $\eta > 0$ and any $k \geq g_{D,N} = O(D^{b_1} N^{b_2})$,

$$\left\| \vartheta_k - \theta_0 \right\|_{\mathbb{R}^D} \lesssim N^{-\eta}.$$

Key proof idea: 'Gradient stability' implies local curvature

Consider the expected negative likelihood

$$\theta \mapsto E_{\theta_0}[-\ell(\theta, Z)] := \frac{1}{2} E_{\theta_0}[(Y - \mathcal{G}(\theta)(X))^2].$$

In 'regular' models, the Hessian satisfies (for $\|v\|_{\mathbb{R}^D} \leq 1$ and some norm $\|\cdot\|_*$)

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

Thus, if there is a **lower bound** for $\|v^T \nabla \mathcal{G}(\theta)\|^2_{L^2(\mathcal{O})}$, then one has *local average curvature*

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

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

- For Schrödinger model, we can verify this via elliptic PDE theory on a neighbourhood of size $\mathrm{diam}(\mathcal{B}) = D^{-4/d}$.

Intermediate result: Globally log-concave Wasserstein approximation

Theorem (Nickl and W, 2020)
With probability $1 - c\exp(-c' N^{d/(2\alpha+d)})$ *under the data:*

(i) *The posterior density* $\pi(\cdot|Z^{(N)})$ *is locally log-concave on* $\mathcal{B}$ *and has a unique mode* $\hat{\theta}_{MAP}$.
(ii) *The proxy density* $\tilde{\pi}(\cdot|Z^{(N)})$ *is globally log-concave with unique mode* $\hat{\theta}_{MAP}$.
(iii) *For all* $N \in \mathbb{N}$, *with* W_2 *denoting Wasserstein distance,*

$$W_2^2(\tilde{\Pi}(\cdot|Z^{(N)}), \Pi(\cdot|Z^{(N)})) \leq \exp(-N^{d/(2\alpha+d)}).$$

- $\tilde{\Pi}(\cdot|Z^{(N)})$ is *not Gaussian* and *non-asymptotic*, thus not based on a Bernstein-von-Mises or Laplace approximation.
- Key property: gradient stability of $\nabla\mathcal{G}$

$$\|v^T \nabla\mathcal{G}(\theta)\|^2_{L^2(\mathcal{O})} \gtrsim D^{-\kappa}\|v\|^2_{\mathbb{R}^D}, \quad \kappa > 0,\ \theta \in \mathcal{B},$$

and regularity of $\mathcal{G}, \nabla\mathcal{G}, \nabla^2\mathcal{G}$.

Gradient stability and **local computability** results have been extended:

- Non-Abelian X-ray transforms [Bohr & Nickl (2021)]
- Darcy flow [Nickl (2022)]
- High-dimensional GLMs [Altmeyer (2022)]

Open questions:

- Numerical investigation
- Polynomial-time initialisation
- Beyond inverse problems?

II. Lower bounds

A specific non-linear regression model

Consider data from random design regression

$$Y_i = \mathcal{G}(\theta)(X_i) + \varepsilon_i, \ \ \varepsilon_i \sim N(0,1).$$

We choose the particular forward operator

$$\mathcal{G}(\theta) = \sqrt{w(\|\theta\|_{\mathbb{R}^D})} \times g(\cdot).$$

- $w : [0, \infty) \to \mathbb{R}$ is a non-decreasing univariate function.
- $g : \mathcal{O} \to [1, 2]$ is an arbitrary smooth, fixed regression function.
- Assume that $D/N \simeq \kappa > 0$.

Prior distribution: $\theta \sim N(0, Id/D)$.

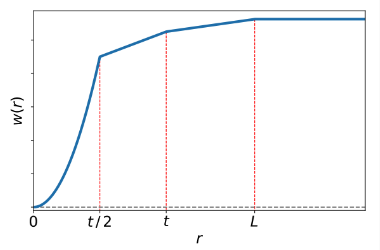


Figure 1: Function w.

Theorem (Bandeira, Maillard, Nickl & W)
Suppose that $\mathcal{G}$ is as constructed before, with appropriate choice of constants. There exists a fixed constant $s \in (0, 1/3)$ such that:

(1) Posterior contraction: *It holds that $\Pi(\{\theta : \|\theta\| \leq s\} | Z^{(N)}) \xrightarrow{N\to\infty} 1$ in probability.*

(2) Unimodality: *The expected likelihood $\ell(\theta)$ is unimodal with mode 0, locally log-concave near 0 and monotonically decreasing in $\|\theta\|_{\mathbb{R}^D}$ on $\mathbb{R}^D$.*

(3) Exponential hitting time: *For any Markov chain $(\vartheta_k : k \geq 1)$ with 'step size' at most $c > 0$, and for some initialisation point $\vartheta_0 \in \{\theta : \|\theta\|_{\mathbb{R}^D} \in (2/3, 2)\}$, the hitting time $\tau_s = \inf_{k\geq 1}\{\|\vartheta_k\| \leq s\}$ is lower bounded with high probability:*

$$\tau_s \geq \exp\big(N/2\big).$$

Lemma: posterior probability ratios of annuli

Let Θ_1, Θ_2 be two disjoint annuli

$$\Theta_1 = \{\theta : \|\theta\| \in [s_1, \eta_1]\}, \quad \Theta_2 = \{\theta : \|\theta\| \in [s_2, \eta_2]\},$$

for $s_1 < \eta_1 < s_2 < \eta_2$.

Proposition (Posterior ratios)

Assume that the prior satisfies, for some constants $\nu > 0$ and $c = c(w) > 0$,

$$\Pi(\Theta_1)/\Pi(\Theta_2) \leq \exp(-N(\nu - c)), \tag{2.1}$$

for some $c > 0$. Suppose that function w is 'slowly increasing' in the region $[s_1, s_2 + \eta_2]$. Then, with high P_0^N-probability as $N \to \infty$,

$$\Pi(\Theta_1 | Z^{(N)})/\Pi(\Theta_2 | Z^{(N)}) \leq \exp(-N\nu).$$

- This holds even when the posterior is concentrating on a $\{\|\theta\| \leq s_1\}$.

A hitting time bound for Markov chains

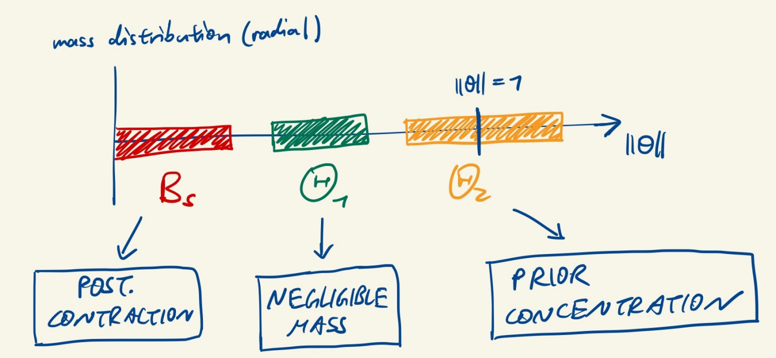
Proposition (cf. Jerrum '03)
Let $(\vartheta_k : k \in \mathbb{N})$ *be any Markov chain with invariant measure* $\mu = \Pi(\cdot|Z^{(N)})$*. Suppose* ϑ_0 *is drawn from the conditional distribution* $\mu(\cdot|\Theta_2)$*. Denote by* τ *the hitting time of the Markov chain onto* Θ_1*. If*

$$\Pi(\Theta_1|Z^{(N)})/\Pi(\Theta_2|Z^{(N)}) \leq \exp(-N\nu),$$

it holds that

$$\Pr(\tau \leq K) \leq Ke^{-N\nu}, \quad K > 0.$$

Intuition for hitting time lower bound



Θ_1 **forms a 'barrier' for (local) Markov chains to go from** Θ_2 **to** B_s**.**

Extensions & Outlook

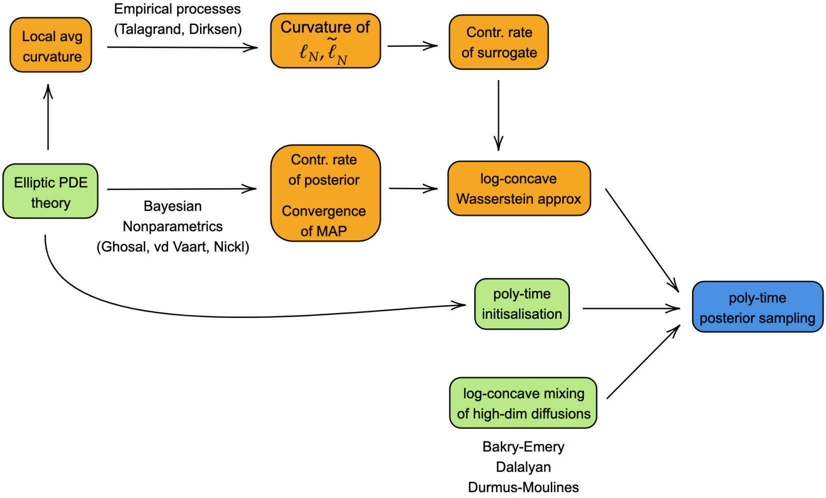
- **'Matern' Priors:** Our results also encompass prior distributions $\Pi = N(0, \Sigma_\alpha)$ with smoothness index $\alpha > d/2$.
- **Lower dimensional models:** Usual choices to optimize 'bias-variance tradeoff' are $D = o(N)$, while we crucially assumed $D \simeq N$.
- **PDE models:** Do free entropy barriers exist in 'real' settings such as PDE models?

References

- **Nickl and W**: On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms. J. Eur. Math. Soc. (2022).
- **Bandeira, Maillard, Nickl and SW**: On free energy barriers in Gaussian priors and failure of MCMC for high-dimensional unimodal distributions. Phil. Trans. R. Soc. A (2023).
- **R Nickl**: Bayesian non-linear statistical inverse problems. EMS lecture notes series (2023).

Thank you for listening!

Proof overview



We now describe the key Langevin-type Markov chain $(\vartheta_k : k \geq 1)$.

Step I: Initialisation. We initialise the algorithm at some specific $\vartheta_0 = \theta_{init}$, where θ_{init} is computable 'in polynomial time'.

Step II: Proxy likelihood construction. Construct a 'proxy' likelihood function $\tilde{\ell}_N$ around θ_{init},

$$\tilde{\ell}_N(\theta) := \alpha(\theta)\ell_N(\theta) - g(\theta).$$

Here $g : \mathbb{R}^D \to \mathbb{R}$ is globally convex, $\alpha : \mathbb{R}^D \to \mathbb{R}$ is a cut-off function. With high probability, we will see that $\tilde{\ell}_N = \ell_N$ locally on

$$\mathcal{B} := \left\{\theta \in \mathbb{R}^D : \left\|\theta - \theta_{0,D}\right\|_{\mathbb{R}^D} \leq \frac{1}{D^{4/d}(\log N)}\right\}.$$

Description of algorithm

The construction from **Step II** induces a proxy posterior distribution,

$$\tilde{\pi}(\theta|Z^{(N)}) \propto e^{\tilde{\ell}_N(\theta)}\pi(\theta), \qquad \log \tilde{\pi}(\theta|Z^{(N)}) = \tilde{\ell}_N(\theta) + \log \pi(\theta) + \text{const}.$$

Step III: Langevin-type Markov chain. For stepsize $\gamma > 0$ and $\xi_k \sim^{i.i.d.} N(0, I_{D\times D})$, define (ϑ_k) with law **P** by

$$\begin{cases} \vartheta_0 & = \theta_{init}, \\ \vartheta_{k+1} & = \vartheta_k + \gamma \nabla \log \tilde{\pi}(\vartheta_k|Z^{(N)}) + \sqrt{2\gamma}\xi_{k+1}. \end{cases}$$

It is the Euler discretisation of the (continuous-time) diffusion process on $\mathbb{R}^D$

$$dL_t = \nabla \log \tilde{\pi}(L_t|Z^{(N)})dt + \sqrt{2}dW_t \qquad t \geq 0,$$

with invariant density $\tilde{\pi}(\cdot|Z^N)$, where W_t is a D-dimensional Brownian motion.