Simultaneous perturbation gradient approximation based Metropolis adjusted Langevin algorithm for inference of ordinary differential equations

Ivan Vujačić*

Mathisca de Gunst*

*Department of Mathematics, Vrije Universiteit Amsterdam, The Netherlands

EYSM 2015, Prague, September 1, 2015

Introduction

Motivation

- System of ordinary differential equations (ODEs) in the standard form

$$\begin{cases} \mathbf{x}'(t) = \mathbf{f}(\mathbf{x}(t), t; \boldsymbol{\theta}), & t \in [0, T], \\ \mathbf{x}(0) = \boldsymbol{\xi}, \end{cases} \tag{1}$$

  where $\mathbf{x}(t), \boldsymbol{\xi} \in \mathbb{R}^d$ and $\boldsymbol{\theta} \in \mathbb{R}^p$.
- $\mathbf{x}(t; \boldsymbol{\theta}, \boldsymbol{\xi})$ denotes the solution of (1) for given $\boldsymbol{\xi}, \boldsymbol{\theta}$.
- Many processes in science and engineering are modelled by (1).

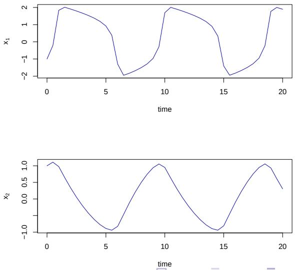
Example: The FitzHugh-Nagumo neural spike potential equations

$$\begin{cases} x_1'(t) = c\{x_1(t) - x_1(t)^3/3 + x_2(t)\}, \\ x_2'(t) = -\frac{1}{c}\{x_1(t) - a + bx_2(t)\}. \end{cases}$$

- x_1 represents the voltage across an axon membrane.
- x_2 summarizes outward currents.

Example:

- $\xi_1 = -1$, $\xi_2 = 1$.
- $a = 0.2$, $b = 0.2$, $c = 3$.



The problem

Noisy observations of $\mathbf{x}(t; \boldsymbol{\theta}_0, \boldsymbol{\xi}_0)$ of some states of the system are available:

$$y_i(t_j) = x_i(t_j; \boldsymbol{\theta}_0, \boldsymbol{\xi}_0) + \varepsilon_i(t_j), \quad i = 1, \ldots, d_1; j = 1, \ldots, n.$$

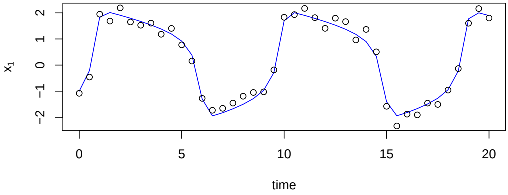
where $0 \leq t_1 \leq \cdots \leq t_n \leq T$.

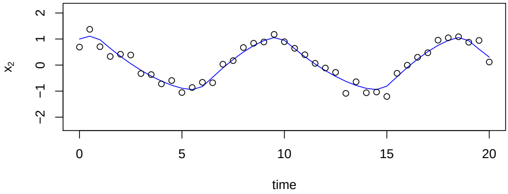
Goal

Estimate $\boldsymbol{\theta}_0$ from the data $\mathbf{Y}$, where $\mathbf{Y} = (y_i(t_j))_{ij}$.

This is inverse problem for the coefficients in a system of ODEs.
If ξ_0 is not known it is considered as parameter and estimated as well.

FhNdata from R package 'CollocInfer'





The system we are interested in

The system, developed by Phillips, that models blood coagulation, where:

- The number of states is large: $d = 83$
- The number of parameters is large: $p = 143$
- Only the first state is observed

Bayesian approach

- π - prior density of $\boldsymbol{\theta}$
- $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_d)$
- $\mathbf{X}(\boldsymbol{\theta}, \boldsymbol{\xi}_0) = (x_i(t_j; \boldsymbol{\theta}, \boldsymbol{\xi}_0))_{ij}$
- $\mathbf{I}_n$ is an identity matrix of order n.
- The posterior density

$$p(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\xi}_0, \boldsymbol{\sigma}) = \pi(\boldsymbol{\theta}) \prod_{j=1}^{d_1} \mathcal{N}\{\mathbf{Y}_{j,\cdot}|\mathbf{X}(\boldsymbol{\theta}, \boldsymbol{\xi}_0)_{j,\cdot}, \sigma_j \mathbf{I}_n\}.$$

The structure of the remainder of the presentation

1. Background
2. The proposed method
3. Numerical results
4. Conclusion

1. Background

MALA (Metropolis Adjusted Langevin Algorithm)

For sampling from $p(\boldsymbol{\theta})$ the MALA proposal is

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^k + \varepsilon^2 \mathbf{M} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^k)/2 + \varepsilon \sqrt{\mathbf{M}} \mathbf{z}^k,$$

where

- $\mathcal{L}(\boldsymbol{\theta}) = \log\{p(\boldsymbol{\theta})\}$
- $\nabla_{\boldsymbol{\theta}}$ - gradient
- $\boldsymbol{\theta}^k$ is the value at k-th step
- $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_p)$
- $\varepsilon > 0$ is the step size
- $\mathbf{M}$ is the weight matrix

The proposal density and acceptance probability are

$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k) = \mathcal{N}(\boldsymbol{\theta}^*|\boldsymbol{\mu}(\boldsymbol{\theta}^k, \varepsilon), \varepsilon^2 \mathbf{M}),$$
$$\alpha = \min\{1, p(\boldsymbol{\theta}^*) q(\boldsymbol{\theta}^k|\boldsymbol{\theta}^*)/p(\boldsymbol{\theta}^k) q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k)\},$$

respectively, where $\boldsymbol{\mu}(\boldsymbol{\theta}^k, \varepsilon) = \boldsymbol{\theta}^k + \varepsilon^2 \mathbf{M} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^k)/2.$

Explicit gradient: Sensitivity equations

To compute the gradient analyticially the sensitivities

$$S_{i,j}(t) = \frac{\mathrm{d}x_i}{\mathrm{d}\theta_j}(t),$$

are required.

$$\begin{cases} S'_{i,j}(t) = \sum_{k=1}^{d} \frac{\partial f}{\partial x_k}(t) S_{k,j}(t) + \frac{\partial f_i}{\partial \theta_j}(t), & t \in [0,T], \\ S_{i,j}(0) = 0. \end{cases}$$

Requires solving a system of *dp* differential equations.

Approximated gradient: Finite difference

The central difference estimate of the j-th partial derivative is

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_j} \approx \frac{\mathcal{L}(\boldsymbol{\theta} + h\mathbf{e}_j) - \mathcal{L}(\boldsymbol{\theta} - h\mathbf{e}_j)}{2h},$$

where

- $\mathbf{e}_j$ is the j-th unit vector
- h is sufficiently small

The central difference estimate requires $2p$ evaluations of $\mathcal{L}$, i.e.

solving d-dimensional system $2p$ times.

SPGA (Simultaneous Perturbation Gradient Approximation)

The two sided (SPGA) is

$$\hat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) = \frac{\mathcal{L}(\boldsymbol{\theta}+h\Delta)-\mathcal{L}(\boldsymbol{\theta}-h\Delta)}{2h}(\Delta_1^{-1},\Delta_2^{-1},\ldots,\Delta_p^{-1})^\top,$$

where

- $\Delta = (\Delta_1,\Delta_2,\ldots,\Delta_p)^\top$ vector of independent Bernoulli random variables
- Δ_k take values -1 and 1 with probability 0.5

Important property

$E\hat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) \to \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}),\ h \to 0.$

3. The proposed method

SPGA-MALA proposal

Substitute the gradient in MALA with its SPGA:

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^k + \varepsilon^2 \mathbf{M} \hat{\nabla}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^k)/2 + \varepsilon \sqrt{\mathbf{M}} \mathbf{z}^k.$$

Acceptance probability needs to be changed accordingly.

Acceptance probability

In view of MALA proposal we require:

$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k,\Delta) = \mathcal{N}(\boldsymbol{\theta}^*|\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}^k,\varepsilon,\Delta),\varepsilon^2\mathbf{M})$$

where

$$\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}^k,\varepsilon,\Delta) = \boldsymbol{\theta}^k + \varepsilon^2\mathbf{M}\hat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}^k)/2.$$

Since Δ can take 2^p values with equal probability it follows

$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k) = \frac{1}{2^p}\sum_{\Delta} q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k,\Delta).$$

The obtained algorithm

Can be viewed as Metropolis Hastings (MH) algorithm:

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^k + \varepsilon^2 \mathbf{M} \hat{\nabla}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^k)/2 + \varepsilon \sqrt{\mathbf{M}} \mathbf{z}^k,$$

$$q(\boldsymbol{\theta}^* | \boldsymbol{\theta}^k) = \frac{1}{2^p} \sum_{\Delta} q(\boldsymbol{\theta}^* | \boldsymbol{\theta}^k, \Delta),$$

$$\alpha = \min\{1, p(\boldsymbol{\theta}^*) q(\boldsymbol{\theta}^k | \boldsymbol{\theta}^*) / p(\boldsymbol{\theta}^k) q(\boldsymbol{\theta}^* | \boldsymbol{\theta}^k)\}.$$

Problem

Evaluating α is intractable for large p.

The modification

Instead of computationally untractable

$$\alpha = \min\{1, p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^k|\boldsymbol{\theta}^*)/p(\boldsymbol{\theta}^k)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k)\},$$

$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k) = \frac{1}{2^p}\sum_{\Delta} q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k, \Delta),$$

use

$$\alpha_\Delta = \min\{1, p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^k|\boldsymbol{\theta}^*, \Delta)/p(\boldsymbol{\theta}^k)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k, \Delta)\}$$

$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k, \Delta) = \mathcal{N}(\boldsymbol{\theta}^*|\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}^k, \varepsilon, \Delta), \varepsilon^2\mathbf{M})$$

where Δ is the drawed (realized) value.

SPGA-MALA Algorithm

1. Generate Δ
2. Compute $\hat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}^k)$.
3. Propose a new value $\boldsymbol{\theta}^* = \boldsymbol{\theta}^k + \varepsilon^2 \mathbf{M}\hat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}^k)/2 + \varepsilon\sqrt{\mathbf{M}}\mathbf{z}^k$.
4. Compute $\alpha_\Delta = \min\{1, p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^k|\boldsymbol{\theta}^*, \Delta)/p(\boldsymbol{\theta}^k)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k)\}$,
5. Generate u from $\mathcal{U}[0,1]$.
6. If $u < \alpha_\Delta$ accept $\boldsymbol{\theta}^*$, otherwise reject.

The algorithm can be viewed as Metropolis-Hastings-Green (MHG) algorithm.

2. Numerical results

Fitz-Hugh Nagumo system

$$\begin{aligned} x_1'(t) &= \theta_3\{x_1(t) - x_1(t)^3/3 + x_2(t)\}, \\ x_2'(t) &= -\frac{1}{\theta_3}\{x_1(t) - \theta_1 + \theta_2 x_2(t)\}. \end{aligned}$$

$\boldsymbol{\theta} = (0.2, 0.2, 3)$ and $\boldsymbol{\xi} = (-1, 1)$.

Fitz-Hugh Nagumo model: results

$d = 2; p = 3$

Sampling method	Time (s)	Mean ESS ($\boldsymbol{\theta}$)	Total time /minimum mean ESS	Relative speed
			$(\theta_1, \theta_2, \theta_3)$	
MALA	363.6	145, 30, 109	12.12	3.4
SPGA-MALA	623.2	84 , 15, 48	41.55	1

Table: Summary of results for 10 runs of the model parameter sampling schemes for Fitz-Hugh Nagumo model with 5000 posterior samples.

α- pinene example

The following model describes the thermal isomerization of α-pinene.

$$
\begin{aligned}
x_1'(t) &= -(\theta_1 + \theta_2)x_1(t), \\
x_2'(t) &= \theta_1 x_1(t), \\
x_3'(t) &= \theta_2 x_1(t) - (\theta_3 + \theta_4)x_3(t) + \theta_5 x_5(t), \\
x_4'(t) &= \theta_3 x_3(t), \\
x_5'(t) &= \theta_4 x_3(t) - \theta_5 x_5(t).
\end{aligned}
$$

$\boldsymbol{\theta} = (0.1, 0.1, 0.3, 0.1, 0.3)$ and $\boldsymbol{\xi} = (1, 0, 0, 0, 0)$.

α-pinene model: results

$d = 5; p = 5$

Sampling method	Time (s)	Mean ESS ($\boldsymbol{\theta}$)	Total time /minimum mean ESS	Relative speed
		$(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$		
MALA	63.7	59, 58, 17, 11, 6	10.62	2.54
SPGA- MALA	134.8	187, 96, 6, 23 , 5	26.96	1

Table: Summary of results for 10 runs of the model parameter sampling schemes for α-pinene example with 5000 posterior samples.

Hockin model of the extrinsic blood coagulation.

$$
\begin{aligned}
x_1'(t) &= -k_{23}x_{28}x_1 - k_{24}x_{34}x_1 - k_{25}x_6x_1 - k_{26}x_3x_1 - k_{27}x_{12}x_1,\\
x_2'(t) &= -k_9x_{28}x_2 - k_{18}x_{31}x_2 - km_{18}x_{33},\\
x_3'(t) &= -k_5x_3x_{20} + k_5x_3x_{20} + k_9x_{28}x_2 - k_{10}x_3x_{21} + k_{10}x_3x_{21} - k_{16}x_3x_{19} +\\
&\quad k_{16}x_3x_{19} + k_{19}x_{34}x_{31} - k_{26}x_3x_1,\\
x_4'(t) &= k_{26}x_3x_1,\\
x_5'(t) &= -k_8x_{12}x_5 - km_8x_{14},\\
x_6'(t) &= kcat_8x_{14} - k_{11}x_{22}x_6 - km_{11}x_8 + k_{14}x_{24}x_{24} + k_{15}x_{24}x_{24} - k_{25}x_6x_1,\\
x_7'(t) &= k_{25}x_6x_1,\\
x_8'(t) &= k_{11}x_{22}x_6 - km_{11}x_8 - k_{12}x_8x_{27} - km_{12}x_9 + kcat_{12}x_9,\\
x_9'(t) &= k_{12}x_8x_{27} - km_{12}x_9 - kcat_{12}x_9,\\
x_{10}'(t) &= -k_1x_{10}x_{20} - km_1x_{11} - k_2x_{10}x_{25} - km_2x_{12},\\
x_{11}'(t) &= k_1x_{10}x_{20} - km_1x_{11},
\end{aligned}
$$

$$
\begin{aligned}
x_{12}'(t) &= k_2 x_{10} x_{25} - km_2 x_{12} - k_3 x_{12} x_{20} + k_3 x_{12} x_{20} - k_6 x_{12} x_{27} - km_6 x_{15} - \\
&\quad k_7 x_{12} x_{28} - km_7 x_{16} - k_8 x_{12} x_5 - km_8 x_{14} + kcat_8 x_{14} - k_{22} x_{12} x_{30} - k_{27} x_{12} x_1, \\
x_{13}'(t) &= k_{27} x_{12} x_1, \\
x_{14}'(t) &= k_8 x_{12} x_5 - km_8 x_{14} - kcat_8 x_{14}, \\
x_{15}'(t) &= k_6 x_{12} x_{27} - km_6 x_{15} - kcat_6 x_{15}, \\
x_{16}'(t) &= kcat_6 x_{15} + k_7 x_{12} x_{28} - km_7 x_{16} - k_{21} x_{16} x_{18} - km_{21} x_{17}, \\
x_{17}'(t) &= k_{21} x_{16} x_{18} - km_{21} x_{17} + k_{22} x_{12} x_{30}, \\
x_{18}'(t) &= -k_{20} x_{28} x_{18} - km_{20} x_{30} - k_{21} x_{16} x_{18} - km_{21} x_{17}, \\
x_{19}'(t) &= -k_{16} x_3 x_{19}, \\
x_{20}'(t) &= -k_1 x_{10} x_{20} - km_1 x_{11} - k_3 x_{12} x_{20} - k_4 x_{28} x_{20} - k_5 x_3 x_{20}, \\
x_{21}'(t) &= -k_{10} x_3 x_{21}, \\
x_{22}'(t) &= k_{10} x_3 x_{21} - k_{11} x_{22} x_6 - km_{11} x_8 + k_{13} x_{24} x_{24} - km_{13} x_{22}, \\
x_{23}'(t) &= -k_{13} x_{24} x_{24} - km_{13} x_{22} - k_{14} x_{24} x_{24} + k_{14} x_{24} x_{24} - k_{15} x_{24} x_{24} + k_{15} x_{24} x_{24},
\end{aligned}
$$

$$x'_{24}(t) = -k_{13}x_{24}x_{24} - km_{13}x_{22} - k_{14}x_{24}x_{24} + k_{14}x_{24}x_{24} - k_{15}x_{24}x_{24} + k_{15}x_{24}x_{24},$$
$$x'_{25}(t) = -k_2x_{10}x_{25} - km_2x_{12} + k_3x_{12}x_{20} + k_4x_{28}x_{20} + k_5x_3x_{20},$$
$$x'_{26}(t) = k_{16}x_3x_{19} - k_{17}x_{28}x_{26} - km_{17}x_{31},$$
$$x'_{27}(t) = -k_6x_{12}x_{27} - km_6x_{15} - k_{12}x_8x_{27} - km_{12}x_9 + k_{14}x_{24}x_{24},$$
$$x'_{28}(t) = -k_4x_{28}x_{20} + k_4x_{28}x_{20} - k_7x_{12}x_{28} - km_7x_{16} - k_9x_{28}x_2 + k_9x_{28}x_2 + kcat_{12}x_9 - k_{17}x_{28}x_{26} - km_{17}x_{31} - k_{20}x_{28}x_{18} - km_{20}x_{30} - k_{23}x_{28}x_1,$$
$$x'_{29}(t) = k_{23}x_{28}x_1,$$
$$x'_{30}(t) = k_{20}x_{28}x_{18} - km_{20}x_{30} - k_{22}x_{12}x_{30},$$
$$x'_{31}(t) = k_{17}x_{28}x_{26} - km_{17}x_{31} - k_{18}x_{31}x_2 - km_{18}x_{33} + kcat_{18}x_{33} - k_{19}x_{34}x_{31} + k_{19}x_{34}x_{31},$$
$$x'_{32}(t) = k_{18}x_{31}x_2 - km_{18}x_{33} - kcat_{18}x_{33},$$
$$x'_{33}(t) = kcat_{18}x_{33} - k_{19}x_{34}x_{31} - k_{24}x_{34}x_1,$$
$$x'_{34}(t) = k_{24}x_{34}x_1.$$

Hockin model: results

$d = 34; p = 43$

Sampling method	Time (s)	Mean ESS ($\boldsymbol{\theta}$)	Total time /minimum mean ESS	Relative speed
		$(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5,$ $\theta_6, \theta_7, \theta_8, \theta_9, \theta_{10})$		
MALA	1.03e+04	5 6 8 7 7 7 6 5 6 8	2060	1
SPGA MALA	180.5	7 6 8 8 7 6 6 5 4 7	45.13	45.65

Table: Summary of results for 10 runs of the model parameter sampling schemes for Hockin model with 5000 posterior samples.

Influenza A virus infection model

$$
\begin{aligned}
x_1'(t) &= -\theta_1 x_1(t) x_3(t),\\
x_2'(t) &= \theta_1 x_1(t) x_3(t) - \theta_2 x_2(t),\\
x_3'(t) &= \theta_3 x_2(t) - \theta_4 x_3(t).
\end{aligned}
$$

$\boldsymbol{\theta} = (0.3, 0.3, 1, 0.5)$ and $\boldsymbol{\xi} = (1, 0.05, 2)$.

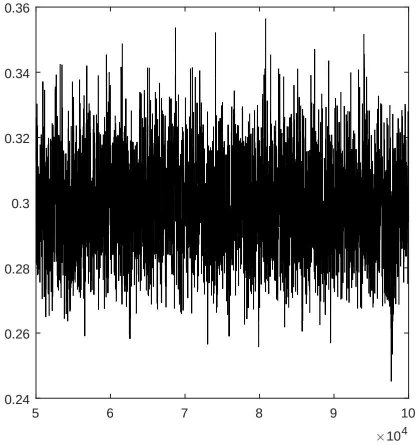


Figure: Traceplot of $\theta_1 = 0.3$; last 50,000 iterations

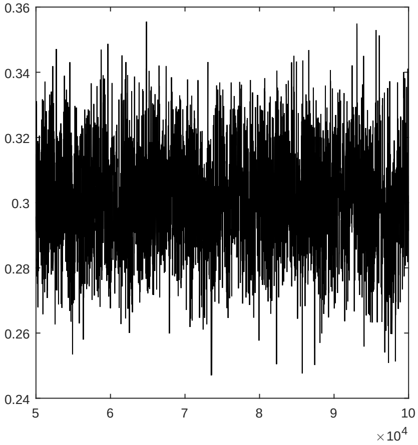


Figure: Traceplot of $\theta_2 = 0.3$; last 50,000 iterations

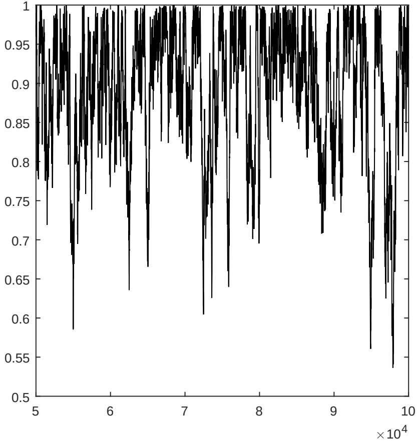


Figure: Traceplot of θ_3=1; last 50,000 iterations

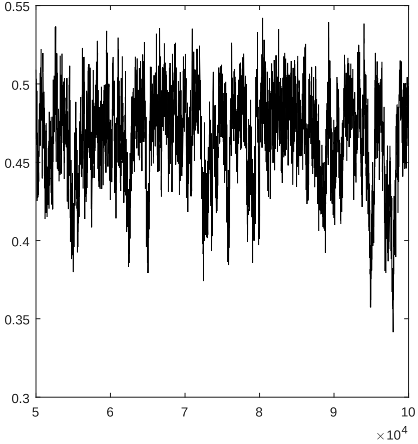


Figure: Traceplot of $\theta_4 = 0.5$; last 50,000 iterations

Conclusion

- SPGA embedded in MALA as a proxy of a gradient.
- Issue of tuning of $\mathbf{M}$ needs to be resolved.

References

Spall, James C. (1992).
Multivariate stochastic approximation using a simultaneous perturbation gradient approximation.
IEEE Transactions on Automatic Control , 37:332–341.

Brooks, S., Gelman, A.,Jones, G. and Meng, Xiao-Li (2011).
Handbook of Markov Chain Monte Carlo.
CRC press .

Girolami, Mark and Calderhead, Ben (2011).
Riemann manifold langevin and hamiltonian monte carlo methods.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73:123–214.