

Miscellaneous

Advanced MCMC Methods for Inverse Problems

Christian Lantuéjoul & Thomas Romary

christian.lantuejoul@mines-paristech.fr

thomas.romary@mines-paristech.fr



Outline

- 1 Bayesian formulation and reminder
- 2 Advanced methods
 - Adaptive MCMC
 - Using the gradient information
 - Annealing approaches
 - Interacting schemes
- 3 Illustration example

The inverse problem

Consider a system \mathcal{V} described by N parameters $X \in \mathcal{X}$

A set of data \mathbf{d}^{obs} has been acquired

$$\mathbf{d}^{\text{obs}} = \mathbf{F}(X) + \epsilon$$

$$X \longrightarrow \mathbf{d}^{\text{obs}}$$

Forward problem

$$\mathbf{d}^{\text{obs}} \longrightarrow X$$

Inverse problem \rightarrow ill-posed

The inverse problem

Common linear(ized) approach

Express iteratively the problem as a linear system and solve:

$$“ X = \mathbf{F}^{-1} \mathbf{d}^{\text{obs}} ”$$

Provides only one “optimal” solution

Bayesian approach [8]

Infer the posterior distribution:

$$\begin{aligned} \pi(X) = \mathbb{P}(X|\mathbf{d}^{\text{obs}}) &= \frac{\mathbb{P}(X)\mathbb{P}(\mathbf{d}^{\text{obs}}|X)}{\mathbb{P}(\mathbf{d}^{\text{obs}})} \\ &\propto \mathbb{P}(X) \exp \left(-\frac{1}{2} \left\| \mathbf{F}(X) - \mathbf{d}^{\text{obs}} \right\|_{C_{\mathbf{d}}^{-1}}^2 \right) \end{aligned}$$

Principle

Build a sequence $\{X_n, n \geq 0\}$ on \mathcal{X} with transition probability P such that π is a *stationary* density for this chain, i.e. $\forall A \in \mathcal{B}(\mathcal{X})$:

$$\int_{\mathcal{X}} P(x, A) \pi(x) dx = \pi(A)$$

Such samples can be used e.g. to compute integrals

$$\pi(h) = \int_{\mathcal{X}} h(x) \pi(x) dx$$

estimating this quantity by

$$S_n(h) = \frac{1}{n} \sum_{i=1}^n h(X_i)$$

for some $h : \mathcal{X} \rightarrow \mathbb{R}$

Metropolis-Hastings samplers

Basic ingredient *proposal distribution* q

Given that the chain is currently at x , a candidate y is accepted with probability

$$\alpha(x, y) = \begin{cases} \min \left\{ 1, \frac{\pi(y)}{\pi(x)} \frac{q(x, y)}{q(y, x)} \right\} & \text{if } \pi(x)q(x, y) > 0 \\ 1 & \text{otherwise} \end{cases}$$

Examples

- 1 the independent sampler (IMH)

$$q(x, y) = q(y)$$

where q is generally the prior in Bayesian inversion

- 2 the symmetric increments random-walk sampler (SIMH)

$$q(x, y) = q(|y - x|)$$

where q can be a zero-mean version of the prior

Typical problems

- Tuning the parameters
- Slowness to converge toward the stationary regime
- Lack of mixing

Outline

- 1 Bayesian formulation and reminder
- 2 Advanced methods
 - Adaptive MCMC
 - Using the gradient information
 - Annealing approaches
 - Interacting schemes
- 3 Illustration example

Adaptive symmetric increments random-walk sampler

Adaptive algorithm of Haario et al. [3] (ASIMH)

y is proposed according to

$$q_{\theta_n}(x, \cdot) = \mathcal{N}(x, \Gamma_n)$$

where $\theta = (\mu, \Gamma)$

Non-decreasing sequence of positive step sizes $\{\gamma_n\}$, such that $\sum_{n=1}^{\infty} \gamma_n = \infty$ and $\sum_{n=1}^{\infty} \gamma_n^{1+\delta} < \infty$ for some $\delta > 0$

In practice, use $\gamma_n = 1/n$

Adaptation procedure

$$\mu_{n+1} = \mu_n + \gamma_{n+1} (X_{n+1} - \mu_n), \quad n \geq 0$$

$$\Gamma_{n+1} = \Gamma_n + \gamma_{n+1} ((X_{n+1} - \mu_n) (X_{n+1} - \mu_n)^t - \Gamma_n)$$

In practice, update is performed every given number of iteration and eventually stopped

Using the gradient information

1 Langevin sampler (LMH)

Assumes π is differentiable on \mathcal{X}

$$q(x, y) \sim \mathcal{N}\left(x + \frac{h^2}{2} \nabla \log(\pi(x)), h^2 I_d\right)$$

2 Hamiltonian Monte-Carlo π is modified so as to include a *kinetic energy* term

A simulation of the Hamiltonian dynamics is performed

Simulated Annealing

M-H Markov chains can be slow to enter their stationary regime

We can write any distribution as a Gibbs distribution

$$\mathbb{P}(\mathbf{b}|\mathbf{d}^{\text{obs}}) \propto e^{-E(\mathbf{b})}$$

Fundamental idea [4]

Control the amplitude of the accepted perturbations thanks to a positive parameter T called the *temperature*.

Simulated Annealing

Principle

We perform a MCMC sampling of the *tempered* posterior distribution

$$\pi_T(\mathbf{b}) = \mathbb{P}(\mathbf{b}|T, \mathbf{d}^{\text{obs}}) \propto e^{-\frac{E(\mathbf{b})}{T}}$$

The temperature T is gradually decreased from T_{\max} to $T = 1$ along the algorithm

At temperature T the acceptance probability is

$$\alpha(\mathbf{b}^l, \mathbf{b}^*) = \min \left(1, e^{\frac{E(\mathbf{b}^l) - E(\mathbf{b}^*)}{T}} \right)$$

Simulated Annealing

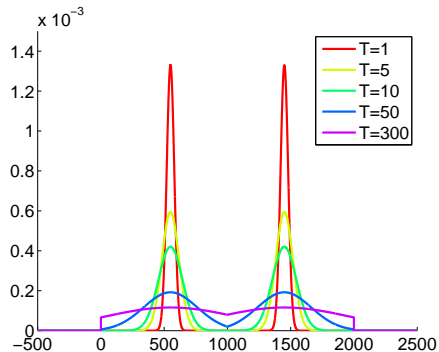


Figure : Effect of the temperature on a bimodal distribution

Simulated Tempering

Let

$$T_0 = 1 < T_1 < \dots < T_K = T_{\max}$$

be a scale of temperatures

Let

$$\pi_i \propto e^{-E(x)/T_i}, i = 0, \dots, K$$

and

$$\mu(X, m) = \sum_{i=0}^K \rho_i \pi_i(X),$$

where $\rho_i > 0$ and $\sum_{i=0}^K \rho_i = 1$

Let also $0 < p_U < 1$ and $0 < p_D < 1$ the probabilities of moving "up", and "down", such that $p_U + p_D < 1$

Simulated Tempering

The principle is to simulate a $\mathcal{X} \times \{0, \dots, K\}$ -valued chain (X_n, M_n)

Let respectively $q_{i \rightarrow i+1}(X_{n+1} | X_n = x_n)$ and $q_{i+1 \rightarrow i}(X_{n+1} | X_n = x_n)$ be the probabilities of transition proposition towards the superior and the inferior temperature level
The acceptance probability of a transition from T_i to T_{i+1} is proportional to

$$\rho_{i \rightarrow i+1}(x_n, x_{n+1}) = \frac{p_D}{p_U} \frac{\pi_{i+1}}{\pi_i} \frac{q_{i \rightarrow i+1}(X_{n+1} = x_{n+1} | X_n = x_n)}{q_{i+1 \rightarrow i}(X_{n+1} = x_{n+1} | X_n = x_n)}$$

Simulated tempering

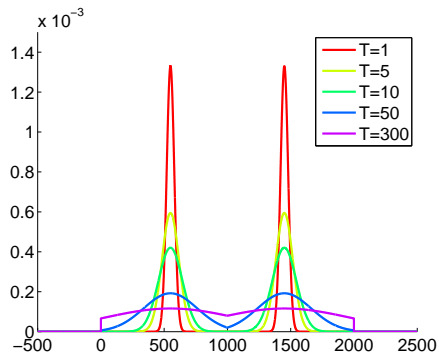


Figure : Effect of the temperature on a bimodal distribution

Interacting schemes

Fundamental idea [2]

To allow several Markov chains at different temperatures to interact in order to sample very efficiently the a posteriori distribution $\mathbb{P}(\mathbf{b}|\mathbf{d}^{\text{obs}})$

The highest temperature chains explore widely the model space while the lowest explore thoroughly around the interesting realizations discovered

Formalization [1]

Simulation from (an approximation of) a nonlinear Markov kernel:

$$P_{\pi}(x, y) = \theta P(x, y) + (1 - \theta) \int P'(z, y) d\pi(z), \quad 0 < \theta < 1$$

where P and P' are two Markov kernels with limit distribution π

Interacting Markov Chains

- $K + 1$ Markov chains runs in parallel
- Each chain l generates samples from π_{T_l} , $l = 0, \dots, K$
- The $(T_l)_{l=0, \dots, K}$ verify:

$$T_0 = 1 < T_1 < \dots < T_K = T_{\max}$$

- *Swaps* between chains at adjacent temperatures are proposed along the algorithm.

see [6, 7, 5, 1] for algorithmic and theoretical details

Outline

- 1 Bayesian formulation and reminder
- 2 Advanced methods
 - Adaptive MCMC
 - Using the gradient information
 - Annealing approaches
 - Interacting schemes
- 3 Illustration example

Illustration example

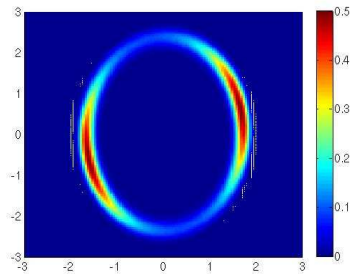
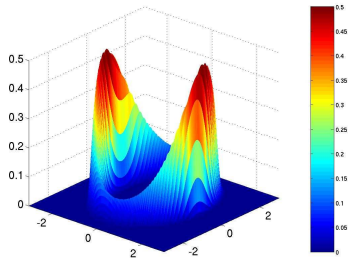
$$\begin{aligned} F : \mathbb{R}^2 &\mapsto \mathbb{R} \\ X = (X_1, X_2) &\rightarrow 2 X_1^2 + X_2^2, \end{aligned} \tag{1}$$

given the following prior on X :

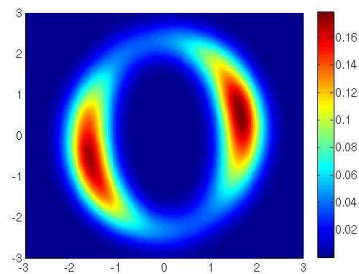
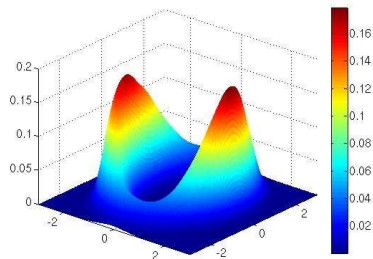
$$X \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}\right) \tag{2}$$

Given a particular realization of X , written $X^* = (1.514, 1.335)$;
we assume that we observe $D^* = F(X^*)$ with an error
 $\varepsilon \sim \mathcal{N}(0, 0.5)$

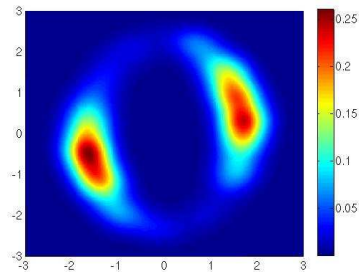
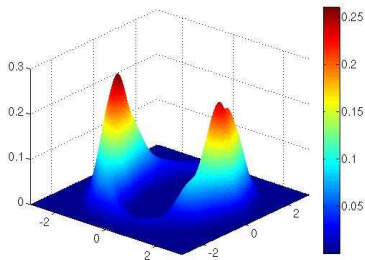
True Posterior



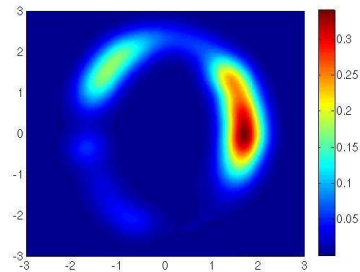
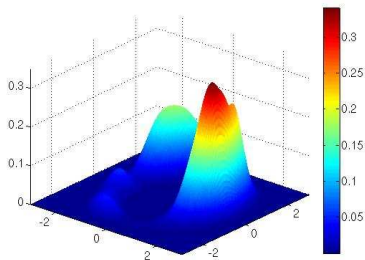
Kernel Density Estimate from an i.i.d. sample



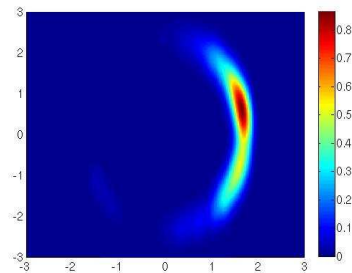
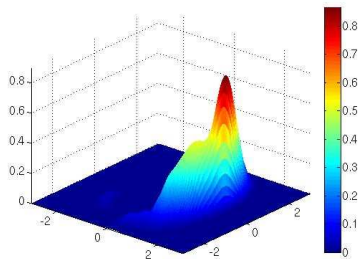
Kernel Density Estimate from the IMH sampler



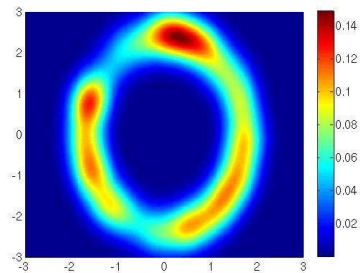
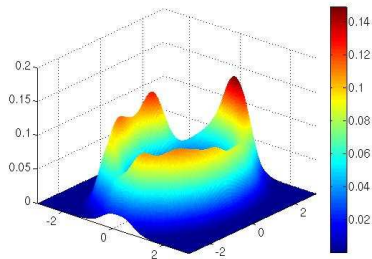
Kernel Density Estimate from the SIMH sampler



Kernel Density Estimate from the Langevin sampler



Kernel Density Estimate from the Adaptive SIMH sampler



Performance criteria

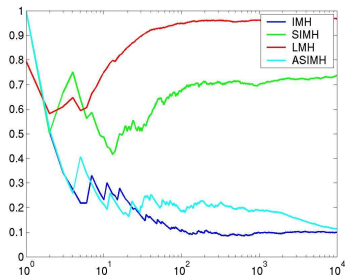


Figure : Empirical acceptance rates, logarithmic scale

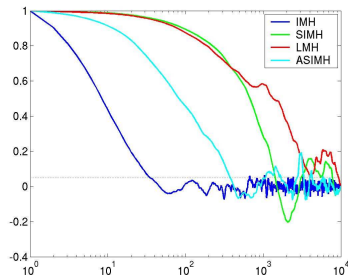
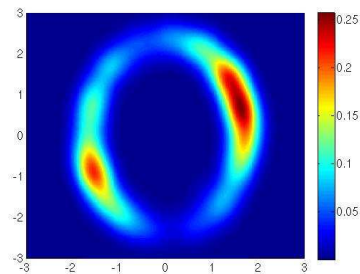
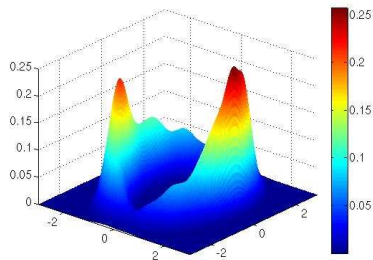


Figure : Autocorrelations along the chain, logarithmic scale

Kernel Density Estimate from one interacting scheme



Performance of one interacting scheme

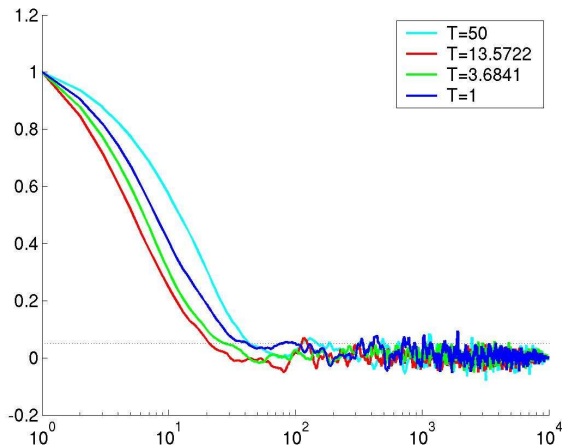


Figure : Autocorrelations along the chains, logarithmic scale

References

- [1] Christophe Andrieu, Ajay Jasra, Arnaud Doucet, and Pierre Del Moral. On nonlinear markov chain monte carlo. *Bernoulli*, 17(3):987–1014, 2011.
- [2] C. J. Geyer. Markov Chain Monte Carlo Maximum Likelihood. In *Computing Science and Statistics: Proceedings of 23rd Symposium on the Interface Interface Foundation*, page 156. American Statistical Association, Fairfax Station, New-York, 1991.
- [3] H. Haario, E. Saksman, and J. Tamminen. An Adaptive Metropolis Algorithm. *Bernoulli*, 7:223–242, 2001.
- [4] N. Metropolis, A. Rosenbluth, M. Rosenbluth, and A. Teller M. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091, 1953.
- [5] T. Romary. History matching of approximated lithofacies models by interacting Markov chains . *Computational Geosciences*, 14(2):343–355, 2009.
- [6] T. Romary. Integrating production data under uncertainty by parallel interacting Markov chains on a reduced dimensional space . *Computational Geosciences*, 13(1):103–122, 2009.
- [7] T. Romary. Bayesian inversion by parallel interacting markov chains. *Inverse Problems in Science and Engineering*, 18(1):111–130, 2010.
- [8] A. Tarantola. *Inverse Problem Theory and Model Parameter Estimation*. SIAM, Philadelphia, 2005.