Behind Markov Chain Monte-Carlo.

E. Moulines

Ecole Nationale Supérieure des Télécommunications

September 24, 2007

Plan

Motivation

Scaling Adaptation

Multidimensional Scaling

Adaptative Metropolis-Hastings Algorithm

An Application

Some theoretical results

Motivation

- MCMC allow to simulate any probability distribution π (typically, large dimensional space)...

Motivation

- MCMC allow to simulate any probability distribution π (typically, large dimensional space)...
- MCMC depends upon tuning parameters, which have a tremendous impact on the sampling performance...

Motivation

- MCMC allow to simulate any probability distribution π (typically, large dimensional space)...
- MCMC depends upon tuning parameters, which have a tremendous impact on the sampling performance...
- Today, Monte-Carlo methods have become a basic tool for inference in complex stochastic models on large datasets.

Motivation

- MCMC allow to simulate any probability distribution π (typically, large dimensional space)...
- MCMC depends upon tuning parameters, which have a tremendous impact on the sampling performance...
- Today, Monte-Carlo methods have become a basic tool for inference in complex stochastic models on large datasets.
- On the top of that, such analysis are often done routinely allowing only limited expert supervision

Motivation

- MCMC allow to simulate any probability distribution π (typically, large dimensional space)...
- MCMC depends upon tuning parameters, which have a tremendous impact on the sampling performance...
- Today, Monte-Carlo methods have become a basic tool for inference in complex stochastic models on large datasets.
- On the top of that, such analysis are often done routinely allowing only limited expert supervision Require to find methods to tune the parameters automatically !

Metropolis-Hastings Algorithm

- Propose a move Y_{n+1} from a transition kernel with density $q(X_n, \cdot)$.

Metropolis-Hastings Algorithm

- Propose a move Y_{n+1} from a transition kernel with density $q(X_n, \cdot)$.
- Accept the move with probability $\alpha(X_n, Y_{n+1})$ where

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}.$$

Metropolis-Hastings Algorithm

- Propose a move Y_{n+1} from a transition kernel with density $q(X_n, \cdot)$.
- Accept the move with probability $\alpha(X_n, Y_{n+1})$ where

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}.$$

- If the move is accepted, set $X_{n+1} = Y_{n+1}$; otherwise, stay at the current position $X_{n+1} = X_n$.

Metropolis Algorithm

- $Y_{k+1} = X_k + Z_{k+1}$ where $Z_{k+1} \sim_{\text{i.i.d.}} q$, and q is symmetric, $q(z) = q(-z)$

Metropolis Algorithm

- $Y_{k+1} = X_k + Z_{k+1}$ where $Z_{k+1} \sim_{\text{i.i.d.}} q$, and q is symmetric, $q(z) = q(-z)$
- In this case, $q(x,y) = q(y,x)$ and the acceptance rate does not depend on the proposal distribution

$$\alpha(x,y) = 1 \wedge \frac{\pi(y)}{\pi(x)}$$

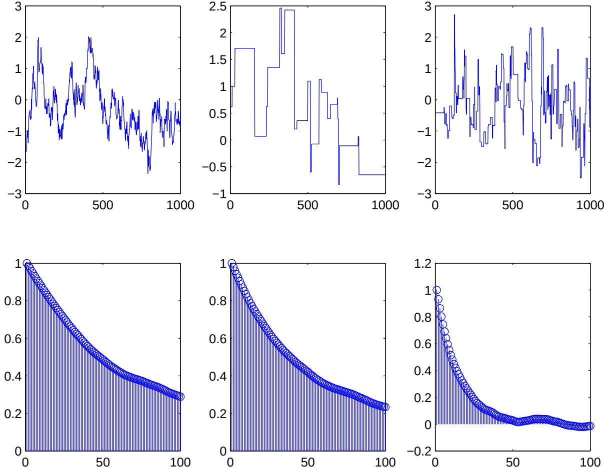
Metropolis Algorithm

- $Y_{k+1} = X_k + Z_{k+1}$ where $Z_{k+1} \sim_{\text{i.i.d.}} q$, and q is symmetric, $q(z) = q(-z)$
- In this case, $q(x, y) = q(y, x)$ and the acceptance rate does not depend on the proposal distribution

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)}{\pi(x)}$$

- ... biased random walk where some moves are rejected.

Scaling


3
2
1
0
−1
−2
−3
0
500
1000
2.5
2
1.5
1
0.5
0
−0.5
−1
0
500
1000
3
2
1
0
−1
−2
−3
0
500
1000
1
0.8
0.6
0.4
0.2
0
0
50
100
1
0.8
0.6
0.4
0.2
0
0
50
100
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
0
50
100

Diffusive Limits

- Stationary distribution: $\pi^{(d)}(x_1, \ldots, x_d) = \prod_{i=1}^{d} f(x_i)$ on $\mathbb{R}^d$ (asymptotic $= d \to \infty$)
- Metropolis proposal: $q_\theta^{(d)}(x_1, \ldots, x_d) \sim \mathcal{N}\left(0, (\theta^2/d)\mathrm{I}_d\right)$... with variance decreasing as $1/d$.

Diffusive Limits

- Stationary distribution: $\pi^{(d)}(x_1, \ldots, x_d) = \prod_{i=1}^{d} f(x_i)$ on $\mathbb{R}^d$ (asymptotic $= d \to \infty$)
- Metropolis proposal: $q_\theta^{(d)}(x_1, \ldots, x_d) \sim \mathcal{N}\left(0, (\theta^2/d)\mathrm{I}_d\right)$... with variance decreasing as $1/d$.
- Interpolated process: $Z_t^{(d)} = X_{[td],1}^{(d)}$... we consider a single component and we speed up the time scale by d.

Diffusive Limits

- Stationary distribution: $\pi^{(d)}(x_1, \ldots, x_d) = \prod_{i=1}^{d} f(x_i)$ on $\mathbb{R}^d$ (asymptotic $= d \to \infty$)
- Metropolis proposal: $q_\theta^{(d)}(x_1, \ldots, x_d) \sim \mathcal{N}\left(0, (\theta^2/d)\mathrm{I}_d\right)$... with variance decreasing as $1/d$.
- Interpolated process: $Z_t^{(d)} = X_{[td],1}^{(d)}$... we consider a single component and we speed up the time scale by d.
- When d becomes large, a single component basically see the mean of the others (**mean-field**)...

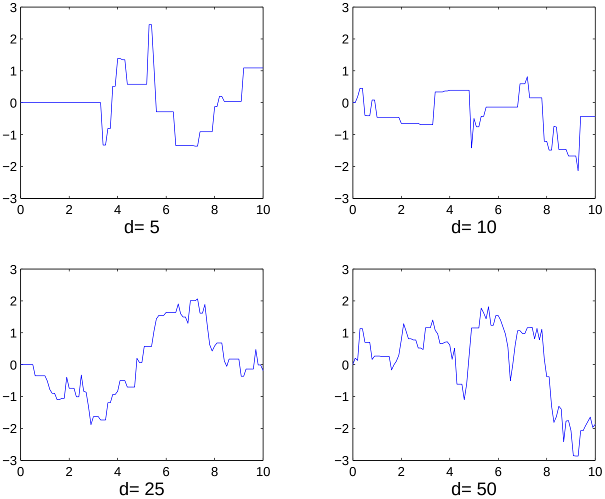


Figure: Diffusive limits for different values of d

Diffusive Limits

- $Z^{(d)} \Rightarrow Z$, where Z solves the Langevin SDE

$$dZ_t = v^{1/2}(\theta)dB_t + (1/2)v(\theta)\nabla \log f(Z_t)dt$$

$$v(\theta) = 2\theta^2 \Phi\left(-\theta\sqrt{I}/2\right)$$

where Φ is the distribution function of $\mathcal{N}(0,1)$ and

Diffusive Limits

- $Z^{(d)} \Rightarrow Z$, where Z solves the Langevin SDE

$$dZ_t = v^{1/2}(\theta) dB_t + (1/2) v(\theta) \nabla \log f(Z_t) dt$$
$$v(\theta) = 2\theta^2 \Phi \left(-\theta \sqrt{I}/2 \right)$$

where Φ is the distribution function of $\mathcal{N}(0,1)$ and I is Fisher Information of the translation model associated to f, $I = \int (f'(x)/f(x))^2 f(x) dx$.

Diffusive Limits

- $Z^{(d)} \Rightarrow Z$, where Z solves the Langevin SDE

$$dZ_t = v^{1/2}(\theta) dB_t + (1/2) v(\theta) \nabla \log f(Z_t) dt$$
$$v(\theta) = 2\theta^2 \Phi\left(-\theta\sqrt{I}/2\right)$$

  where Φ is the distribution function of $\mathcal{N}(0,1)$ and I is Fisher Information of the translation model associated to f, $I = \int (f'(x)/f(x))^2 f(x) dx$.
- $v(\theta)$ is the speed of the diffusion: $Z_t = \tilde{Z}_{v(\theta)t}$ where $\{\tilde{Z}_t\}$ is a solution of the normalized Langevin SDE

$$d\tilde{Z}_t = dB_t + (1/2)\nabla \log f(\tilde{Z}_t) dt.$$

Speed / Acceptance rate

- Mean Acceptance rate (stationary regime)

$$\tau^{(d)}(\theta) = \iint \pi^{(d)}(\mathbf{x}) q_\theta^{(d)}(\mathbf{y}-\mathbf{x}) \left\{ 1 \wedge \frac{\pi^{(d)}(\mathbf{y})}{\pi^{(d)}(\mathbf{x})} \right\} d\mathbf{x} d\mathbf{y} \ .$$

Speed / Acceptance rate

- Mean Acceptance rate (stationary regime)

$$\tau^{(d)}(\theta) = \iint \pi^{(d)}(\mathbf{x}) q_\theta^{(d)}(\mathbf{y}-\mathbf{x}) \left\{ 1 \wedge \frac{\pi^{(d)}(\mathbf{y})}{\pi^{(d)}(\mathbf{x})} \right\} d\mathbf{x} d\mathbf{y} \; .$$

- Result: $\tau^{(\infty)}(\theta) = \lim_{d\to\infty} \tau^{(d)}(\theta)$ exists and it is possible to relate the speed of the diffusion to the mean acceptance rate !

$$v(\theta) = \tau^{(\infty)}(\theta) \left\{ \Phi^{-1}(\tau^{(\infty)}(\theta)/2) \right\}^2$$

Speed / Acceptance rate

- Mean Acceptance rate (stationary regime)

$$\tau^{(d)}(\theta) = \iint \pi^{(d)}(\mathbf{x}) q_\theta^{(d)}(\mathbf{y}-\mathbf{x}) \left\{ 1 \wedge \frac{\pi^{(d)}(\mathbf{y})}{\pi^{(d)}(\mathbf{x})} \right\} d\mathbf{x} d\mathbf{y} \, .$$

- Result: $\tau^{(\infty)}(\theta) = \lim_{d\to\infty} \tau^{(d)}(\theta)$ exists and it is possible to relate the speed of the diffusion to the mean acceptance rate !

$$v(\theta) = \tau^{(\infty)}(\theta) \left\{ \Phi^{-1}(\tau^{(\infty)}(\theta)/2) \right\}^2$$

- The speed is optimal for the value θ_* of the parameter which satisfies $\tau^{(\infty)}(\theta_*) = \bar{\tau} \approx 0.234$...

How to control the Acceptance Rate

- Objective: Finding the scaling factor θ solving

$$h(\theta) \stackrel{\text{def}}{=} \iint \alpha(\mathbf{x}, \mathbf{y}) q_\theta(\mathbf{y} - \mathbf{x}) \pi(\mathbf{x}) d\mathbf{x} d\mathbf{y} - \bar{\tau} = 0,$$

where $\alpha(\mathbf{x}, \mathbf{y}) = \{1 \wedge \pi(\mathbf{y}) / \pi(\mathbf{x})\}$.

How to control the Acceptance Rate

- Objective: Finding the scaling factor θ solving

$$h(\theta) \stackrel{\text{def}}{=} \iint \alpha(\mathbf{x},\mathbf{y}) q_\theta(\mathbf{y}-\mathbf{x}) \pi(\mathbf{x}) d\mathbf{x} d\mathbf{y} - \bar{\tau} = 0,$$

  where $\alpha(\mathbf{x},\mathbf{y}) = \{1 \wedge \pi(\mathbf{y})/\pi(\mathbf{x})\}$.
- Under general assumptions, $\theta \to h(\theta)$ is monotone with $\lim_{\theta\to 0^+} h(\theta) = 1 - \bar{\tau} > 0$ and $\lim_{\theta\to\infty} h(\theta) = -\bar{\tau} < 0$...

How to control the Acceptance Rate

- **Objective:** Finding the scaling factor θ solving

$$h(\theta) \stackrel{\text{def}}{=} \iint \alpha(\mathbf{x},\mathbf{y}) q_\theta(\mathbf{y}-\mathbf{x}) \pi(\mathbf{x}) d\mathbf{x} d\mathbf{y} - \bar{\tau} = 0,$$

  where $\alpha(\mathbf{x},\mathbf{y}) = \{1 \wedge \pi(\mathbf{y})/\pi(\mathbf{x})\}$.

- Under general assumptions, $\theta \to h(\theta)$ is monotone with $\lim_{\theta\to 0^+} h(\theta) = 1 - \bar{\tau} > 0$ and $\lim_{\theta\to\infty} h(\theta) = -\bar{\tau} < 0$... But $h(\theta)$ cannot be computed explicitly !

How to control the Acceptance Rate

- Objective: Finding the scaling factor θ solving

$$h(\theta) \stackrel{\text{def}}{=} \iint \alpha(\mathbf{x}, \mathbf{y}) q_\theta(\mathbf{y} - \mathbf{x}) \pi(\mathbf{x}) d\mathbf{x} d\mathbf{y} - \bar{\tau} = 0,$$

  where $\alpha(\mathbf{x}, \mathbf{y}) = \{1 \wedge \pi(\mathbf{y})/\pi(\mathbf{x})\}$.
- Under general assumptions, $\theta \to h(\theta)$ is monotone with $\lim_{\theta \to 0^+} h(\theta) = 1 - \bar{\tau} > 0$ and $\lim_{\theta \to \infty} h(\theta) = -\bar{\tau} < 0$... But $h(\theta)$ cannot be computed explicitly !
- Nevertheless, denoting θ_k the scaling value at iteration k, $\alpha(X_k, Y_{k+1}) - \bar{\tau}$ may be seen as a "noisy" observation of $h(\theta_k)$...

How to control the Acceptance Rate

- Objective: Finding the scaling factor θ solving

$$h(\theta) \stackrel{\text{def}}{=} \iint \alpha(\mathbf{x},\mathbf{y}) q_\theta(\mathbf{y}-\mathbf{x})\pi(\mathbf{x}) d\mathbf{x} d\mathbf{y} - \bar{\tau} = 0,$$

  where $\alpha(\mathbf{x},\mathbf{y}) = \{1 \wedge \pi(\mathbf{y})/\pi(\mathbf{x})\}$.

- Under general assumptions, $\theta \to h(\theta)$ is monotone with $\lim_{\theta \to 0^+} h(\theta) = 1 - \bar{\tau} > 0$ and $\lim_{\theta \to \infty} h(\theta) = -\bar{\tau} < 0$... But $h(\theta)$ cannot be computed explicitly !
- Nevertheless, denoting θ_k the scaling value at iteration k, $\alpha(X_k, Y_{k+1}) - \bar{\tau}$ may be seen as a "noisy" observation of $h(\theta_k)$...
- **Suggest to use a stochastic approximation procedure to tune θ.**

Controlled Metropolis Algorithm

- Proposition & Accept/Reject

$$Y_{k+1} = X_k + \theta_k \mathcal{N}(0, \mathrm{Id})$$

$$X_{k+1} = \begin{cases} Y_{k+1} & \text{with prob. } \alpha(X_k, Y_{k+1}) \\ X_k & \text{otherwise} \end{cases}$$

Controlled Metropolis Algorithm

- Proposition & Accept/Reject

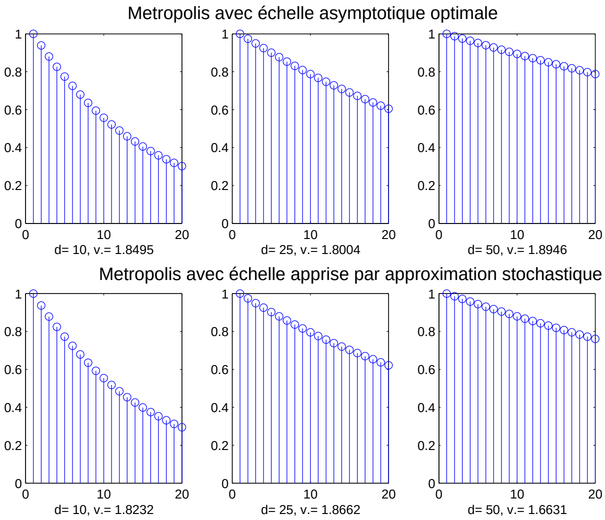
$$Y_{k+1} = X_k + \theta_k \mathcal{N}(0, \mathrm{Id})$$

$$X_{k+1} = \begin{cases} Y_{k+1} & \text{with prob. } \alpha(X_k, Y_{k+1}) \\ X_k & \text{otherwise} \end{cases}$$

- Update the scaling factor

$$\theta_{k+1} = \theta_k + \gamma_{k+1} \left\{ \alpha(X_k, Y_{k+1}) - \bar{\tau} \right\}$$

where $\lim_{k\to\infty} \gamma_k = 0$ and $\sum_{k=1}^{\infty} \gamma_k = \infty$.


Metropolis avec échelle asymptotique optimale
d= 10, v.= 1.8495
d= 25, v.= 1.8004
d= 50, v.= 1.8946
Metropolis avec échelle apprise par approximation stochastique
d= 10, v.= 1.8232
d= 25, v.= 1.8662
d= 50, v.= 1.6631

Multidimensional scaling

- Same asymptotic analysis ($d \to \infty$) with

$$\pi^{(d)}_{\Sigma_d}(\mathbf{x}) = |\Sigma_d|^{-1}\pi^{(d)}\left(\Sigma_d^{-1}\mathbf{x}\right), \quad \pi^{(d)}(x_1,\ldots,x_d) = \prod_{i=1}^{d} f(x_i)$$

$$q \sim N(0, (\sigma^2/d)\mathrm{Id})$$

then $Z_t^{(d)} = X_{[td],1}$ converges to the solution a Langevin SDE.

Multidimensional scaling

- Same asymptotic analysis ($d \to \infty$) with

$$\pi^{(d)}_{\Sigma_d}(\mathbf{x}) = |\Sigma_d|^{-1}\pi^{(d)}\left(\Sigma_d^{-1}\mathbf{x}\right), \quad \pi^{(d)}(x_1, \dots, x_d) = \prod_{i=1}^{d} f(x_i)$$

$$q \sim N(0, (\sigma^2/d)\mathrm{Id})$$

  then $Z_t^{(d)} = X_{[td],1}$ converges to the solution a Langevin SDE.
- the target acceptance rate (0.234...) which maximizes the speed of the limiting diffusion is independent from Σ_d, but the achievable maximal speed is strongly affected by Σ_d...

Multidimensional scaling

- Same asymptotic analysis ($d \to \infty$) with

$$\pi^{(d)}_{\Sigma_d}(\mathbf{x}) = |\Sigma_d|^{-1}\pi^{(d)}\left(\Sigma_d^{-1}\mathbf{x}\right), \quad \pi^{(d)}(x_1, \ldots, x_d) = \prod_{i=1}^{d} f(x_i)$$

$$q \sim N(0, (\sigma^2/d)\mathrm{Id})$$

then $Z^{(d)}_t = X_{[td],1}$ converges to the solution a Langevin SDE.

- the target acceptance rate (0.234...) which maximizes the speed of the limiting diffusion is independent from Σ_d, but the achievable maximal speed is strongly affected by Σ_d... loss

$$\lim_d \frac{d^{-1}\sum_{i=1}^{d}\lambda_{d,i}^2}{\left(d^{-1}\sum_{i=1}^{d}\lambda_{d,i}\right)^2}$$

where $\lambda_{d,i}$ eigenvalues of Σ_d.

Adaptive MCMC with multidim. scaling

1. Simulate

$$Y_{k+1} = X_k + \mathcal{N}(0, \sigma_k \Gamma_\mathrm{k})$$

$$X_{k+1} = \begin{cases} Y_{k+1} & \text{with proba. } \alpha(X_k, Y_{k+1}) \\ X_k & \text{otherwise} \end{cases}$$

Adaptive MCMC with multidim. scaling

1. Simulate

$$Y_{k+1} = X_k + \mathcal{N}(0, \sigma_k \Gamma_\mathrm{k})$$
$$X_{k+1} = \begin{cases} Y_{k+1} & \text{with proba. } \alpha(X_k, Y_{k+1}) \\ X_k & \text{otherwise} \end{cases}$$

2. Update the target mean and covariance

$$\mu_{k+1} = \mu_k + \gamma_{k+1}(X_{k+1} - \mu_k)$$
$$\Gamma_{k+1} = \Gamma_k + \gamma_{k+1}\left\{(X_{k+1} - \mu_k)(X_{k+1} - \mu_k)^\mathtt{T} - \Gamma_k\right\}$$

Adaptive MCMC with multidim. scaling

1. Simulate

$$Y_{k+1} = X_k + \mathcal{N}(0, \sigma_k \Gamma_{\mathrm{k}})$$
$$X_{k+1} = \begin{cases} Y_{k+1} & \text{with proba. } \alpha(X_k, Y_{k+1}) \\ X_k & \text{otherwise} \end{cases}$$

2. Update the target mean and covariance

$$\mu_{k+1} = \mu_k + \gamma_{k+1}(X_{k+1} - \mu_k)$$
$$\Gamma_{k+1} = \Gamma_k + \gamma_{k+1}\left\{(X_{k+1} - \mu_k)(X_{k+1} - \mu_k)^{\mathtt{T}} - \Gamma_k\right\}$$

3. Control the global scale of the proposal

$$\sigma_{k+1} = \sigma_k + \gamma_{k+1}\left(\alpha(X_k, Y_{k+1}) - \bar{\tau}\right)$$

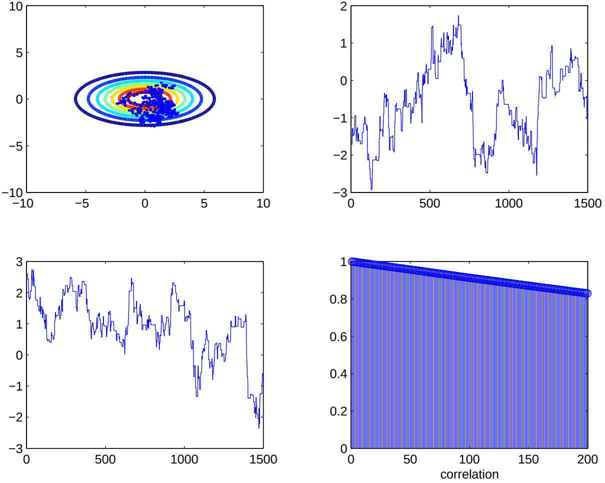


Figure: $d = 12,\ \pi \sim \mathcal{N}(0, \Gamma),\ \mathrm{cond}(\Gamma) \approx 100,\ q \sim \mathcal{N}(0, (2.32^2/d)\,\mathrm{I})$

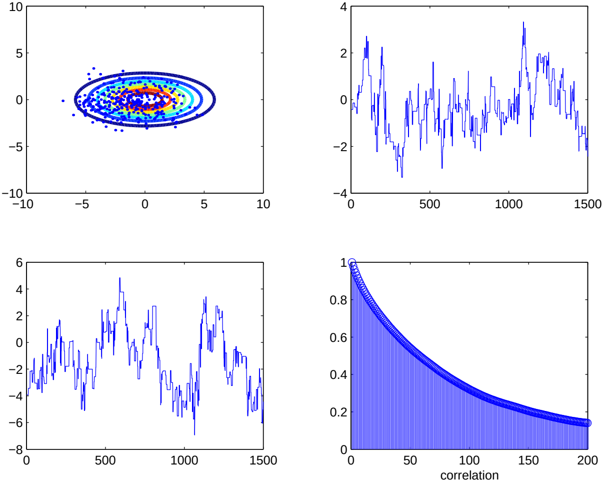


Figure: $d = 12$, $\pi \sim \mathcal{N}(0, \Gamma)$, $\mathrm{cond}(\Gamma) \approx 100$, $q \sim \mathcal{N}(0, 2.32^2/d\Gamma)$

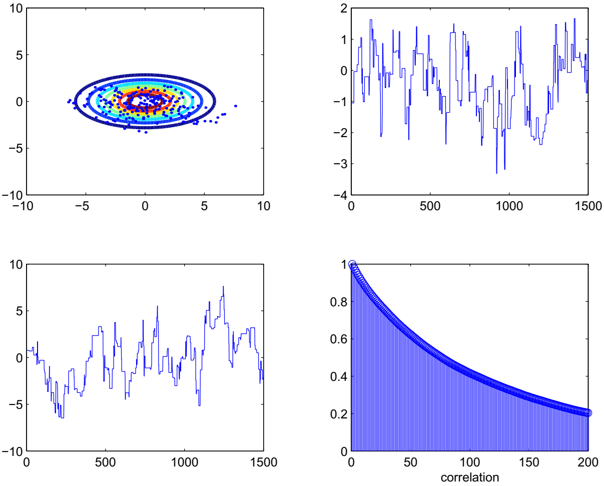


Figure: $d = 12$, $\pi \sim \mathcal{N}(0, \Gamma)$, $\mathrm{cond}(\Gamma) \approx 100$, $q \sim \mathcal{N}(0, \sigma_k \Gamma_k)$, with adaptive multidimensional scaling

Tricks and Improvements

- No need to estimate the covariance matrix at each iteration [batch means = OK]

Tricks and Improvements

- No need to estimate the covariance matrix at each iteration [batch means = OK]
- Update the eigendecomposition of the covariance matrix directly [Oja and the many improvements since then].

Tricks and Improvements

- No need to estimate the covariance matrix at each iteration [batch means = OK]
- Update the eigendecomposition of the covariance matrix directly [Oja and the many improvements since then].
- In large dimension, it is often ore sensible to use hybrid algorithm, to update a subset of the parameters... the eigendecomposition can help there to find the directions which are worthwhile to update.

Tricks and Improvements

- No need to estimate the covariance matrix at each iteration [batch means = OK]
- Update the eigendecomposition of the covariance matrix directly [Oja and the many improvements since then].
- In large dimension, it is often ore sensible to use hybrid algorithm, to update a subset of the parameters... the eigendecomposition can help there to find the directions which are worthwhile to update.
- In presence of non-linear correlation π, estimating a single covariance matrix is not enough. In this case, non-linear ACP methods (e.g. locally linear) are better suited...

Metropolis-Hastings with independent proposals

- Propose Y_{k+1} from a pdf q independently from the past

Metropolis-Hastings with independent proposals

- Propose Y_{k+1} from a pdf q independently from the past
- Accept the move with prob. $\alpha(X_k, Y_{k+1})$, where

$$\alpha(x,y) = 1 \wedge \frac{\pi(y)q(x)}{\pi(x)q(y)}$$

Metropolis-Hastings with independent proposals

- Propose Y_{k+1} from a pdf q independently from the past
- Accept the move with prob. $\alpha(X_k, Y_{k+1})$, where

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)q(x)}{\pi(x)q(y)}$$

- Geometrically ergodic if $\pi(x) \leq Mq(x)$ and the rate is controlled by $1/M$ (similar to accept/reject).

Metropolis-Hastings with independent proposals

- Propose Y_{k+1} from a pdf q independently from the past
- Accept the move with prob. $\alpha(X_k, Y_{k+1})$, where

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)q(x)}{\pi(x)q(y)}$$

- Geometrically ergodic if $\pi(x) \leq Mq(x)$ and the rate is controlled by $1/M$ (similar to accept/reject).
- Similar to the A/R algorithm, efficient if the proposal q is close to π...

Metropolis-Hastings with independent proposals

- Idea: Choose the proposal distribution in a parametric family $(q_\theta, \theta \in \Theta)$.

Metropolis-Hastings with independent proposals

- Idea: Choose the proposal distribution in a parametric family $(q_\theta, \theta \in \Theta)$.
- Example: mixture of Gaussians
  1. easy to sample
  2. universal approximation

Metropolis-Hastings with independent proposals

- Idea: Choose the proposal distribution in a parametric family $(q_\theta, \theta \in \Theta)$.
- Example: mixture of Gaussians
  1. easy to sample
  2. universal approximation
- Objective: on-line adaptation of the parameter by minimizing the Kullback divergence

$$\mathrm{KL}(\pi \| q_\theta) = \int \log\left(\frac{\pi(x)}{q_\theta(x)}\right) \pi(x) dx \; .$$

Metropolis-Hastings with independent proposals

- Idea: Choose the proposal distribution in a parametric family $(q_\theta, \theta \in \Theta)$.
- Example: mixture of Gaussians
  1. easy to sample
  2. universal approximation
- Objective: on-line adaptation of the parameter by minimizing the Kullback divergence

$$\mathrm{KL}(\pi \| q_\theta) = \int \log(\frac{\pi(x)}{q_\theta(x)})\pi(x)dx \ .$$

- Method: On-line EM algorithm (see ICASSP 2006)

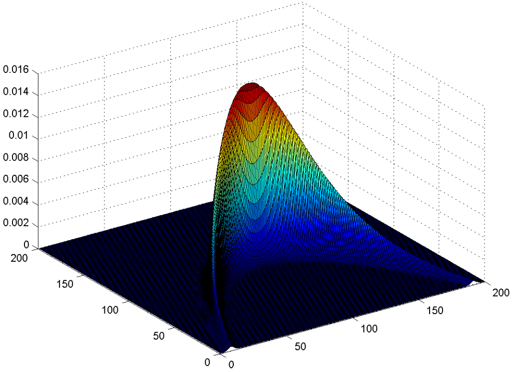
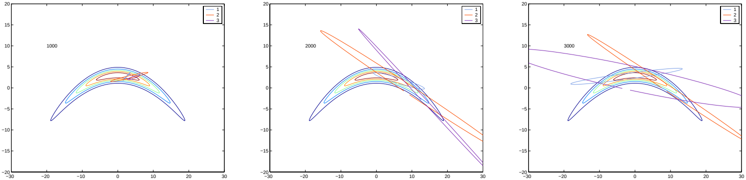
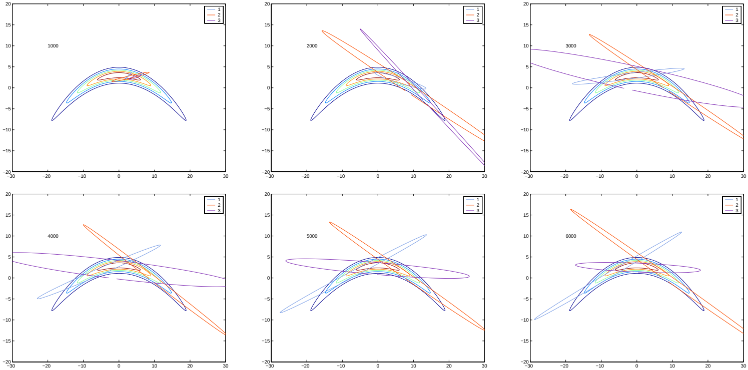
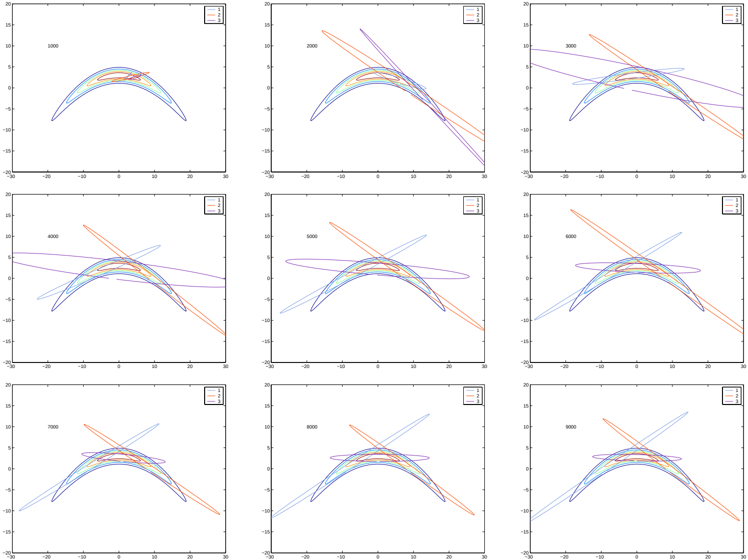


Figure: Banana shaped target distribution


1000
2000
3000
1
2
3


1000
2000
3000
4000
5000
6000
1
2
3


1000
2000
3000
4000
5000
6000
7000
8000
9000
1
2
3

Results (Andrieu & Moulines, 2006)

- Law of Large Numbers (under assumptions that do not imply the cvge of θ_k)

$$n^{-1} \sum_{k=1}^{n} [f(X_k) - \pi(f)] \xrightarrow{\text{a.s}}_{\bar{\mathrm{P}}_\star} 0 \; .$$

*asymptotically, adaptation cost

Results (Andrieu & Moulines, 2006)

- Law of Large Numbers (under assumptions that do not imply the cvge of θ_k)

$$n^{-1} \sum_{k=1}^{n} [f(X_k) - \pi(f)] \xrightarrow{\text{a.s}}_{\bar{\mathrm{P}}_\star} 0 \; .$$

- Central Limit Theorem (if $\lim_k \theta_k$ exists)

$$n^{-1/2} \sum_{k=1}^{n} [f(X_k) - \pi(f)] \xrightarrow{\mathcal{D}}_{\bar{\mathrm{P}}_\star} Z \; ,$$

with Z characteristic function $\bar{\mathrm{E}}_\star \left[\exp(-\frac{1}{2}\sigma^2(\theta_\infty, f)t^2)\right]$ and $\sigma^2(\theta_\infty, f)$ variance of the MCMC under $\theta_\infty{}^*$

*asymptotically, adaptation cost

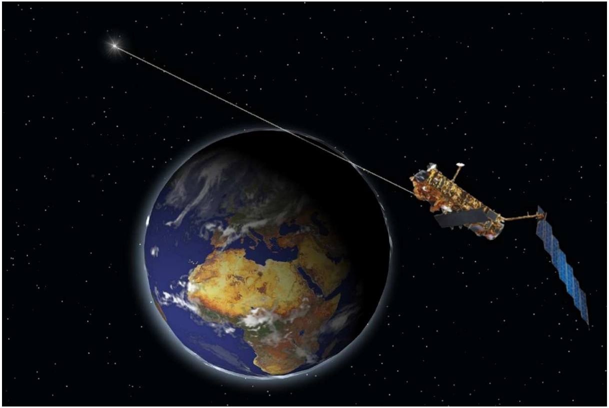


Figure: Global monitoring of gaseous matters (ozone layer) and aerosol concentrations by occultation of stars

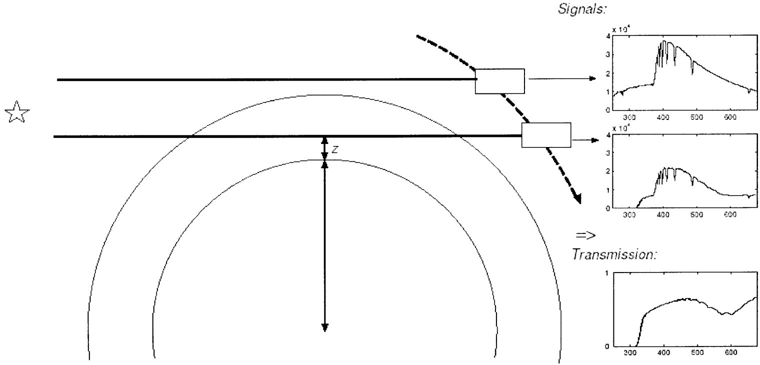


Figure: Principle of the measurement of the transmittance spectrum

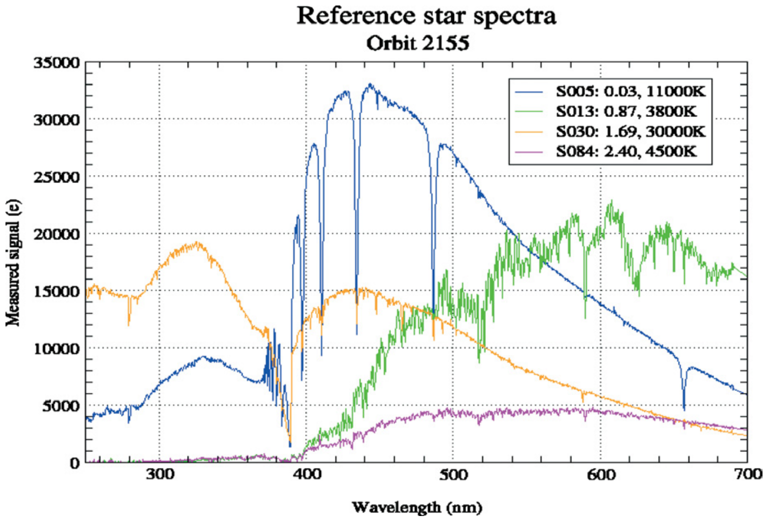


Figure: Spectrum of the star for considered wavelengths

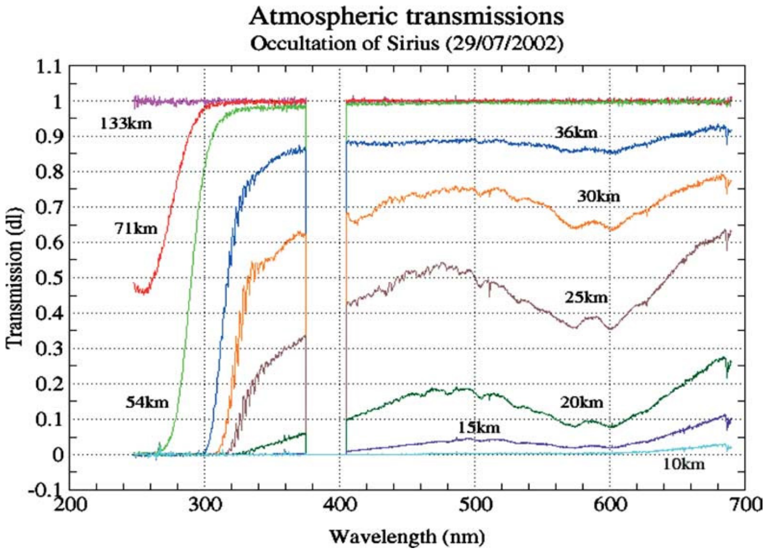


Figure: Atmospheric transmittance at different tangential altitude (height). Sirius

Model

- **Principle:** $T(\lambda, z) = \exp\left(-\sum_g \alpha_g(\lambda) N_g(z)\right)$ (Beer & Lambert)
  1. $T(\lambda, z)$ transmittance at λ and tangential altitude z

Model

- **Principle:** $T(\lambda, z) = \exp\left(-\sum_g \alpha_g(\lambda) N_g(z)\right)$ (Beer & Lambert)
  1. $T(\lambda, z)$ transmittance at λ and tangential altitude z
  2. $N_g(z)(\mathrm{mol/cm^2})$ integrated quantity of gaseous matter (O_3, H_2O, NO_2 ...) at tangential height z. Related to the concentration $z \mapsto \rho_g(z)$ by

$$N_g(z) = \int_{\ell(z)} \rho_g[z(s)]ds, \quad \ell(z) = \text{line of sight}$$

Model

- **Principle:** $T(\lambda, z) = \exp\left(-\sum_g \alpha_g(\lambda) N_g(z)\right)$ (Beer & Lambert)
  1. $T(\lambda, z)$ transmittance at λ and tangential altitude z
  2. $N_g(z)(\mathrm{mol/cm^2})$ integrated quantity of gaseous matter (O_3, H_2O, NO_2 ...) at tangential height z. Related to the concentration $z \mapsto \rho_g(z)$ by

$$N_g(z) = \int_{\ell(z)} \rho_g[z(s)]ds, \quad \ell(z) = \text{line of sight}$$

  3. $\alpha_g(\lambda)$ absorption coefficient of gaseous species g at frequency λ.

- Altitude discretization (approx. 1 km) and $\rho_g(z)$ assumed constant for altitude diff. less than the step-size:

$$N_g(z_i) = \sum_{j=1}^{J} \ell_{i,j} R_{g,j} \qquad R_{g,j} = \rho_g(z_j)$$

- Altitude discretization (approx. 1 km) and $\rho_g(z)$ assumed constant for altitude diff. less than the step-size:

$$N_g(z_i) = \sum_{j=1}^{J} \ell_{i,j} R_{g,j} \qquad R_{g,j} = \rho_g(z_j)$$

- Prior model for the concentration: Gaussian Linear State Space Model, i.e. $R_{g,j} = [01] X_{g,j}$

$$X_{g,j} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} X_{g,j-1} + \begin{pmatrix} \sigma \\ 0 \end{pmatrix} \mathcal{N}(0,1)$$

Measurements

- Measurements: noisy estimates of the transmittance at frequencies $\lambda_1, \ldots, \lambda_I$ et d'altitudes $z_1, \ldots, z_J$

$$y(\lambda_i, z_j) = T(\lambda_i, z_j) + \varepsilon(\lambda_i, z_j) \ ,$$

where $\varepsilon(\lambda_i, z_j)$ measurement noise (independent, Gaussian, known variance)...

Measurements

- Measurements: noisy estimates of the transmittance at frequencies $\lambda_1, \dots, \lambda_I$ et d'altitudes $z_1, \dots, z_J$

$$y(\lambda_i, z_j) = T(\lambda_i, z_j) + \varepsilon(\lambda_i, z_j) \ ,$$

where $\varepsilon(\lambda_i, z_j)$ measurement noise (independent, Gaussian, known variance)...

- Objective: Infer the posterior distribution of the gaseous component concentration $\{R_{g,j}, j = 1, \dots, J, g = 1, \dots, G\}$... Well-posed Non-Linear Inverse Problem!

Main Characteristics

- Huge number of measurements: $I \approx 1500$ fréquencies, $J \approx 100$ height: 150000 measurement for a single occultation experiments (and up to 10 occultation experiment / day)...

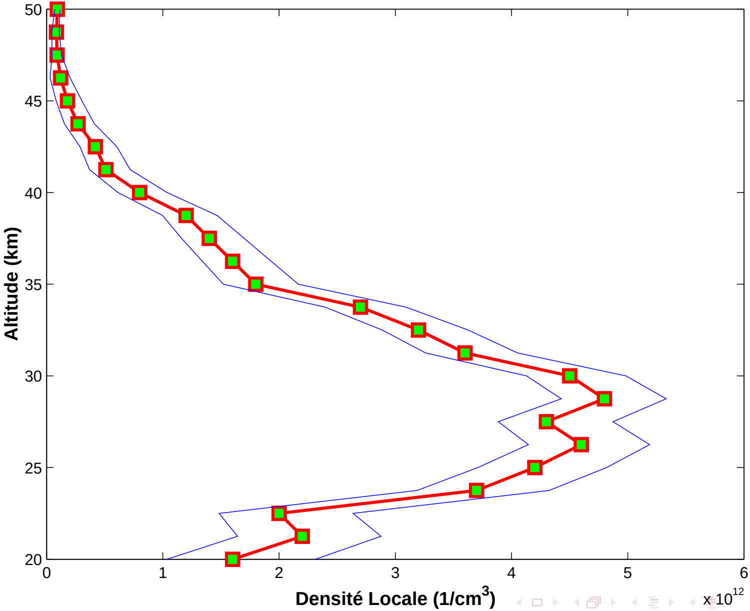
Main Characteristics

- Huge number of measurements: $I \approx 1500$ fréquencies, $J \approx 100$ height: 150000 measurement for a single occultation experiments (and up to 10 occultation experiment / day)...
- Huge number of variables $J \times G \approx 500$.

Main Characteristics

- Huge number of measurements: $I \approx 1500$ fréquencies, $J \approx 100$ height: 150000 measurement for a single occultation experiments (and up to 10 occultation experiment / day)...
- Huge number of variables $J \times G \approx 500$.
- Variability of the experimental set-up star emission spectrum, atmospheric turbulence, line of sight ...

Adaptation is vital !


Altitude (km)
50
45
40
35
30
25
20
0
1
2
3
4
5
6
Densité Locale (1/cm3)
x 10^12

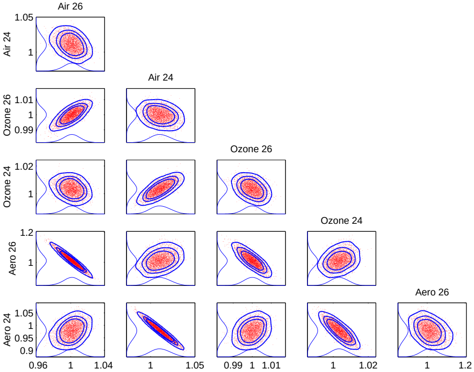


Figure: Joint and marginal distributions of two gaseous components at 24 and 26 km

Conclusions

- Adaptive MCMC methods are a new class of simulation strategy, which is likely to help the dissemination of these techniques at large.

Conclusions

- Adaptive MCMC methods are a new class of simulation strategy, which is likely to help the dissemination of these techniques at large.
- There are many possible ways to adapt a simulation strategy.

Conclusions

- Adaptive MCMC methods are a new class of simulation strategy, which is likely to help the dissemination of these techniques at large.
- There are many possible ways to adapt a simulation strategy. Most often, it is more difficult to find appropriate adaptation criteria rather than to design the on-line procedure itself.

Conclusions

- Adaptive MCMC methods are a new class of simulation strategy, which is likely to help the dissemination of these techniques at large.
- There are many possible ways to adapt a simulation strategy. Most often, it is more difficult to find appropriate adaptation criteria rather than to design the on-line procedure itself.
- Sensible criterion

Conclusions

- Adaptive MCMC methods are a new class of simulation strategy, which is likely to help the dissemination of these techniques at large.
- There are many possible ways to adapt a simulation strategy. Most often, it is more difficult to find appropriate adaptation criteria rather than to design the on-line procedure itself.
- Sensible criterion $\leftrightarrow$ understand the chain dynamic (simulation bottleneck)

Conclusions

- Adaptive MCMC methods are a new class of simulation strategy, which is likely to help the dissemination of these techniques at large.
- There are many possible ways to adapt a simulation strategy. Most often, it is more difficult to find appropriate adaptation criteria rather than to design the on-line procedure itself.
- Sensible criterion $\leftrightarrow$ understand the chain dynamic (simulation bottleneck) $\leftrightarrow$ asymptotic analysis (dimension, fluid limit, etc.)

Conclusions

There have been many advances since the first works on this subject in the end of the 90's

Conclusions

There have been many advances since the first works on this subject in the end of the 90's nevertheless, there are a lot of problems left:

Conclusions

There have been many advances since the first works on this subject in the end of the 90's nevertheless, there are a lot of problems left:

1. Hybrid Algorithms.

Conclusions

There have been many advances since the first works on this subject in the end of the 90's nevertheless, there are a lot of problems left:

1. Hybrid Algorithms.
2. Links with controlled Markov chain (policy) and reinforcement learning.

Conclusions

There have been many advances since the first works on this subject in the end of the 90's nevertheless, there are a lot of problems left:

1. Hybrid Algorithms.
2. Links with controlled Markov chain (policy) and reinforcement learning.
3. Coupling MCMC (serial) and particle (parallel) methods.

Conclusions

There have been many advances since the first works on this subject in the end of the 90's nevertheless, there are a lot of problems left:

1. Hybrid Algorithms.
2. Links with controlled Markov chain (policy) and reinforcement learning.
3. Coupling MCMC (serial) and particle (parallel) methods.
4. Extensions to trans-dimensionnal simulation methods

Conclusions

There have been many advances since the first works on this subject in the end of the 90's nevertheless, there are a lot of problems left:

1. Hybrid Algorithms.
2. Links with controlled Markov chain (policy) and reinforcement learning.
3. Coupling MCMC (serial) and particle (parallel) methods.
4. Extensions to trans-dimensionnal simulation methods
5. most needed ToolBox (AdapBUGS !)

Ingredients

- $(P_\theta, \theta \in \Theta)$ a family of transition kernels with target distribution π.

Ingredients

- $(P_\theta, \theta \in \Theta)$ a family of transition kernels with target distribution π.
- $h : \Theta \to \Theta$ the objective estimating function; the optimal parameters are the roots of the non-linear equation $h(\theta) = 0$ (Z-estimator).

Ingredients

- $(P_\theta, \theta \in \Theta)$ a family of transition kernels with target distribution π.
- $h : \Theta \to \Theta$ the objective estimating function; the optimal parameters are the roots of the non-linear equation $h(\theta) = 0$ (Z-estimator).
- $H : \Theta \times \mathsf{X} \to \Theta$ an estimating function: for all $\theta \in \Theta$,

$$h(\theta) \stackrel{\text{def}}{=} \iint_{\mathsf{X}} H(x, \theta) \pi(dx) \ .$$

Ingredients

- $(P_\theta, \theta \in \Theta)$ a family of transition kernels with target distribution π.
- $h : \Theta \to \Theta$ the objective estimating function; the optimal parameters are the roots of the non-linear equation $h(\theta) = 0$ (Z-estimator).
- $H : \Theta \times \mathsf{X} \to \Theta$ an estimating function: for all $\theta \in \Theta$,

$$h(\theta) \stackrel{\text{def}}{=} \iint_{\mathsf{X}} H(x, \theta)\pi(dx) \ .$$

- Algorithm:

$$X_{k+1} \sim P_{\theta_k}(X_k, \cdot)$$
$$\theta_{k+1} = \theta_k + \gamma_{k+1} H(\theta_k, X_{k+1})$$

Problems and Questions.

$\{(X_k, \theta_k)\}$ is a non-homogeneous Markov Chain but... $\{X_k\}$ is not a Markov Chain ! Q: Is it still ergodic ?

Problems and Questions.

$\{(X_k, \theta_k)\}$ is a non-homogeneous Markov Chain but... $\{X_k\}$ is not a Markov Chain ! Q: Is it still ergodic ?

1. Limit Theorems for Additive Functionals

$$n^{-\gamma} \sum_{k=1}^{n} \left(\psi_{\theta_k}(X_k) - \int_{\mathsf{X}} \psi_{\theta_k}(x) \pi(dx) \right)$$

Problems and Questions.

$\{(X_k, \theta_k)\}$ is a non-homogeneous Markov Chain but... $\{X_k\}$ is not a Markov Chain ! Q: Is it still ergodic ?

1. Limit Theorems for Additive Functionals

$$n^{-\gamma} \sum_{k=1}^{n} \left(\psi_{\theta_k}(X_k) - \int_{\mathsf{X}} \psi_{\theta_k}(x)\pi(dx) \right)$$

2. Rate of Convergence (?)

$$\| \mathrm{E}_{(x,\theta)}[f(X_k)] - \pi(f) \|_{\mathrm{TV}} \leq C \|f\|_\infty r(k)$$

Ergodicity is not Automatically preserved...

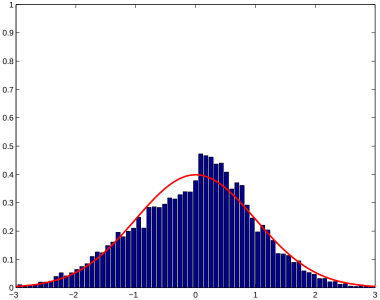


Figure: Metropolis algorithm on $\mathbb{R}$. Target $\pi = \mathcal{N}(0,1)$, Proposal $q = \mathcal{N}(0,\theta^2)$. Adaptation: $\theta^2 = \theta_+^2$ if $X_k \geq 0$ et $\theta^2 = \theta_-^2$ if $X_k < 0$.

Assumptions: Geometric Ergodicity

There exists a function $V : \mathsf{X} \to [1, \infty]$ and a set C such that for all $\mathsf{K} \subset \Theta$ compact,

- Foster-Lyapunov: $\sup_{\theta \in \mathsf{K}} P_\theta V \leq \lambda_{\mathsf{K}} V + b_{sfK} \mathbb{1}_{\mathsf{C}}$

† Coupling + Sharpening the Lindvall inequality

Assumptions: Geometric Ergodicity

There exists a function $V : \mathsf{X} \to [1, \infty]$ and a set C such that for all $\mathsf{K} \subset \Theta$ compact,

- Foster-Lyapunov: $\sup_{\theta \in \mathsf{K}} P_\theta V \leq \lambda_{\mathsf{K}} V + b_{sfK} \mathbb{1}_{\mathsf{C}}$
- Minorization: $\inf_{x \in \mathsf{C}} \inf_{\theta \in \mathsf{K}} P_\theta(x, \cdot) \geq \delta_{\mathsf{K}} \nu(\cdot)$

† Coupling + Sharpening the Lindvall inequality

Assumptions: Geometric Ergodicity

There exists a function $V : \mathsf{X} \to [1, \infty]$ and a set C such that for all $\mathsf{K} \subset \Theta$ compact,

- Foster-Lyapunov: $\sup_{\theta \in \mathsf{K}} P_\theta V \leq \lambda_{\mathsf{K}} V + b_{sfK} \mathbb{1}_{\mathsf{C}}$
- Minorization: $\inf_{x \in \mathsf{C}} \inf_{\theta \in \mathsf{K}} P_\theta(x, \cdot) \geq \delta_{\mathsf{K}} \nu(\cdot)$

(Douc & M.,2003)† There exist constants $\rho_{\mathsf{K}} < 1$ and $C_{\mathsf{K}} < \infty$, depending explicitly on λ_{K}, b_{K} and δ_{K}, such that

$$\sup_{\theta \in \mathcal{K}} \| P_\theta^n f - \pi(f) \|_V \leq C_{\mathsf{K}} \|f\|_V \rho_{\mathsf{K}}^n \quad \|f\|_V = \sup |f(x)|/V(x)$$

†Coupling + Sharpening the Lindvall inequality

Assumptions: Smoothness

For all $\mathsf{K} \subset \Theta$ compact, $(\theta, \theta') \in \mathsf{K} \times \mathsf{K}$

- $\|P_\theta f - P_{\theta'} f\|_V \leq C_{\mathsf{K}} \, \|f\|_V \, |\theta - \theta'|$

Assumptions: Smoothness

For all $\mathsf{K} \subset \Theta$ compact, $(\theta, \theta') \in \mathsf{K} \times \mathsf{K}$

- $\|P_\theta f - P_{\theta'} f\|_V \leq C_{\mathsf{K}} \, \|f\|_V \, |\theta - \theta'|$
- $\|H(\theta, \cdot) - H(\theta', \cdot)\|_{V^{1/2}} \leq C_{\mathsf{K}} |\theta - \theta'|$

Assumptions: Smoothness

For all $\mathsf{K} \subset \Theta$ compact, $(\theta, \theta') \in \mathsf{K} \times \mathsf{K}$

- $\|P_\theta f - P_{\theta'} f\|_V \leq C_\mathsf{K} \, \|f\|_V \, |\theta - \theta'|$
- $\|H(\theta, \cdot) - H(\theta', \cdot)\|_{V^{1/2}} \leq C_\mathsf{K} |\theta - \theta'|$

(Andrieu & M.,2005) Existence of a solution $\hat{f}_\theta$ to the Poisson equation $f - \pi(f) = \hat{f}_\theta - P_\theta \hat{f}_\theta$;

Assumptions: Smoothness

For all $\mathsf{K} \subset \Theta$ compact, $(\theta, \theta') \in \mathsf{K} \times \mathsf{K}$

- $\|P_\theta f - P_{\theta'} f\|_V \leq C_{\mathsf{K}} \, \|f\|_V \, |\theta - \theta'|$
- $\|H(\theta, \cdot) - H(\theta', \cdot)\|_{V^{1/2}} \leq C_{\mathsf{K}} |\theta - \theta'|$

(Andrieu & M.,2005) Existence of a solution $\hat{f}_\theta$ to the Poisson equation $f - \pi(f) = \hat{f}_\theta - P_\theta \hat{f}_\theta$; $\hat{f}_\theta$ is Lipshitz $\|\hat{f}_\theta - \hat{f}_{\theta'}\|_V \leq C_{\mathsf{K}} |\theta - \theta'|$ for all $(\theta, \theta') \in \mathsf{K} \times \mathsf{K}$.

Assumptions: Smoothness

For all $\mathsf{K} \subset \Theta$ compact, $(\theta, \theta') \in \mathsf{K} \times \mathsf{K}$

- $\|P_\theta f - P_{\theta'} f\|_V \leq C_{\mathsf{K}} \, \|f\|_V \, |\theta - \theta'|$
- $\|H(\theta, \cdot) - H(\theta', \cdot)\|_{V^{1/2}} \leq C_{\mathsf{K}} |\theta - \theta'|$

(Andrieu & M.,2005) Existence of a solution $\hat{f}_\theta$ to the Poisson equation $f - \pi(f) = \hat{f}_\theta - P_\theta \hat{f}_\theta$; $\hat{f}_\theta$ is Lipshitz $\|\hat{f}_\theta - \hat{f}_{\theta'}\|_V \leq C_{\mathsf{K}} |\theta - \theta'|$ for all $(\theta, \theta') \in \mathsf{K} \times \mathsf{K}$. Satisfied by most Metropolis-Hastings algorithms, (but not always easy).

Error decomposition

Existence of solutions to Poisson Eqs. $\Rightarrow$

$$f(X_k) - \pi(f) = \hat{f}_{\theta_k}(X_k) - P_{\theta_k}\hat{f}_{\theta_k}(X_k),$$

Error decomposition

Existence of solutions to Poisson Eqs. $\Rightarrow$

$$f(X_k) - \pi(f) = \hat{f}_{\theta_k}(X_k) - P_{\theta_k}\hat{f}_{\theta_k}(X_k),$$

Error Decomposition

$$\hat{f}_{\theta_k}(X_k) - P_{\theta_k}\hat{f}_{\theta_k}(X_k) = \left(\hat{f}_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}}\hat{f}_{\theta_{k-1}}(X_{k-1})\right) + \left(\hat{f}_{\theta_k}(X_k) - \hat{f}_{\theta_{k-1}}(X_k)\right) + \left(P_{\theta_{k-1}}\hat{f}_{\theta_{k-1}}(X_{k-1}) - P_{\theta_k}\hat{f}_{\theta_k}(X_k)\right)$$

First term: martingale. Second term: Lipshitz (cvgce de θ_k not necessary for LLN). Third term: disappear in the summations...

Error decomposition

Existence of solutions to Poisson Eqs. $\Rightarrow$

$$f(X_k) - \pi(f) = \hat{f}_{\theta_k}(X_k) - P_{\theta_k}\hat{f}_{\theta_k}(X_k),$$

Error Decomposition

$$\hat{f}_{\theta_k}(X_k) - P_{\theta_k}\hat{f}_{\theta_k}(X_k) = \left(\hat{f}_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}}\hat{f}_{\theta_{k-1}}(X_{k-1})\right) + \left(\hat{f}_{\theta_k}(X_k) - \hat{f}_{\theta_{k-1}}(X_k)\right) + \left(P_{\theta_{k-1}}\hat{f}_{\theta_{k-1}}(X_{k-1}) - P_{\theta_k}\hat{f}_{\theta_k}(X_k)\right)$$

First term: martingale. Second term: Lipshitz (cvgce de θ_k not necessary for LLN). Third term: disappear in the summations...

Assumptions: Stability of the adaptation procedure and convergence

Lyapunov function $w : \Theta \to [0, \infty]$ such that

1. Level-sets $\mathcal{W}_M \stackrel{\text{def}}{=} \{\theta \in \Theta, w(\theta) \leq M\} \subset \Theta$ are compact,

Assumptions: Stability of the adaptation procedure and convergence

Lyapunov function $w : \Theta \to [0, \infty]$ such that

1. Level-sets $\mathcal{W}_M \stackrel{\text{def}}{=} \{\theta \in \Theta, w(\theta) \leq M\} \subset \Theta$ are compact,
2. $\langle \nabla w(\theta), h(\theta) \rangle \leq 0$

Assumptions: Stability of the adaptation procedure and convergence

Lyapunov function $w : \Theta \to [0, \infty]$ such that

1. Level-sets $\mathcal{W}_M \stackrel{\text{def}}{=} \{\theta \in \Theta, w(\theta) \leq M\} \subset \Theta$ are compact,
2. $\langle \nabla w(\theta), h(\theta) \rangle \leq 0$
3. the closure of $w\left(\{\theta \in \Theta, \langle \nabla w(\theta), h(\theta) \rangle = 0\}\right)$ has an empty interior

Assumptions: Stability of the adaptation procedure and convergence

Lyapunov function $w : \Theta \to [0, \infty]$ such that

1. Level-sets $\mathcal{W}_M \stackrel{\text{def}}{=} \{\theta \in \Theta, w(\theta) \leq M\} \subset \Theta$ are compact,
2. $\langle \nabla w(\theta), h(\theta) \rangle \leq 0$
3. the closure of $w\left(\{\theta \in \Theta, \langle \nabla w(\theta), h(\theta) \rangle = 0\}\right)$ has an empty interior

(Andrieu & M. et Priouret, 2005) convergence of stochastic approximation $d(\theta_k, \mathcal{L}) \to 0$ p.s. under verifiable assumptions h

Stochastic Approximation: an introduction

- Let Θ be the domain of allowable values for a vector of parameters θ.
- Two fundamental problems of interest:

  Problem 1. Find the value(s) of a vector $\theta \in \Theta$ that minimize a scalar-valued loss function $w(\theta)$

Stochastic Approximation: an introduction

- Let Θ be the domain of allowable values for a vector of parameters θ.
- Two fundamental problems of interest:

  Problem 1. Find the value(s) of a vector $\theta \in \Theta$ that minimize a scalar-valued loss function $w(\theta)$

  Problem 2. Find the value(s) of $\theta \in \Theta$ that solve the equation $h(\theta) = 0$ for some vector-valued function h. Frequently (but not necessarily) $h(\theta) = \nabla w(\theta)$

Stochastic root-finding problem

- Focus is on finding θ (i.e., $\theta^\star$) such that $h(\theta^\star) = 0$ where $h(\theta)$ is typically a nonlinear function of θ assuming that only noisy measurements of $h(\theta)$ are available

$$\theta_{k+1} = \theta_k + \gamma_{k+1} Y_{k+1} \quad Y_{k+1} = h(\theta_k) + \text{"noise"}$$

- Above problem arises frequently in practice
  - Optimization with noisy measurements ($h(\theta)$ represents gradient of loss function)
  - Machine learning
  - ... and adaptive MCMC.

back

Existence solutions to the Poisson Equation

- For any compact subset $\mathcal{K} \subset \Theta$ and for any $r \in [0,1]$ there exist constants C and $\rho < 1$ such that for all $\psi \in \mathcal{L}_{V^r}$ and all $\theta \in \mathcal{K}$
$$\|P_\theta^k \psi - \pi(\psi)\|_{V^r} \leq C\rho^k \, \|\psi\|_{V^r}.$$
- Therefore, for all $\theta, x \in \Theta \times \mathsf{X}$ and $\psi \in \mathcal{L}_{V^r}$,
$$\sum_{k=0}^{\infty} |P_\theta^k \psi(x) - \pi(\psi)| < \infty$$
and
$$u \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} (P_\theta^k \psi - \pi(\psi))$$
is a solution of Poisson's equation: $u - P_\theta u = \psi - \pi(\psi)$.

Regularity of the Solutions to the Poisson Equation

for any function f and k,

$$
\begin{aligned}
P_\theta^k f - P_{\theta'}^k f &= \sum_{j=1}^{k-1} P_\theta^j \left(P_\theta - P_{\theta'}\right) P_{\theta'}^{k-j-1} f \\
&= \sum_{j=1}^{k-1} P_\theta^j \left(P_\theta - P_{\theta'}\right) \left(P_{\theta'}^{k-j-1} f - \pi(f)\right) \\
&= \sum_{j=1}^{k-1} \left(P_\theta^j - \pi\right) \left(P_\theta - P_{\theta'}\right) \left(P_{\theta'}^{k-j-1} f - \pi(f)\right)
\end{aligned}
$$

Regularity of the Solutions to the Poisson Equation

- The geometric ergodicity and the regularity of the transition kernel implies that

$$
\begin{aligned}
|P_\theta^k f - P_{\theta'}^k f| &\leq C_{\theta,\theta'} \sum_{j=1}^{k-1} \rho_\theta^j \rho_{\theta'}^{k-j} |\theta - \theta'| \\
&\leq C_{\theta,\theta'} \rho^k |\theta - \theta'| \quad \rho = \rho_\theta \wedge \rho_{\theta'} \ .
\end{aligned}
$$

- Denote by $\hat{f}_\theta$ the solution of the Poisson equation $\hat{f}_\theta - P_\theta \hat{f}_\theta = f - \pi(f)$. Then

$$
\hat{f}_\theta - \hat{f}_{\theta'} = \sum_{k=1}^{\infty} (P_\theta^k f - P_{\theta'}^k f)
$$

and the regularity follows from the preceding bound.