Adaptive and Interacting Monte Carlo methods for Bayesian analysis

Gersende Fort

LTCI, CNRS & TELECOM ParisTech
Paris, France

Slides available:
`http://perso.telecom-paristech.fr/∼gfort/Communications.html`

Outline

Introduction

Markov chain Monte Carlo

Importance Sampling

Conclusion

Bibliography

Outline

- Introduction
  - Bayesian Analysis
  - Monte Carlo methods
- Markov chain Monte Carlo
  - Algorithm
  - Convergence of the method
  - Accuracy of the approximation
  - Proposal distribution and efficiency
  - Adaptive MCMC
  - Interacting methods
- Importance Sampling
  - Algorithm
  - Convergence of the method
  - Monitoring the convergence
  - Adaptive Importance sampling
  - Curse of dimensionality
- Conclusion
  - MCMC vs Importance Sampling
  - Burn In
  - Parallelization
- Bibliography

Statistical model

Learning about *parameters* through observations:

- a *likelihood* of the observations $\mathbf{y}$ given some parameters of interest $\mathbf{x}$

$$p(\mathbf{y}|\mathbf{x})$$

- a *prior* on the parameters of interest

$$p(\mathbf{x})$$

- yielding the a posteriori distribution of the parameters

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{\int p(\mathbf{y}|\mathbf{x}')p(\mathbf{x}')d\mathbf{x}'}$$

Statistical model

Learning about *parameters* through observations:

- a *likelihood* of the observations $\mathbf{y}$ given some parameters of interest $\mathbf{x}$

$$p(\mathbf{y}|\mathbf{x})$$

- a *prior* on the parameters of interest

$$p(\mathbf{x})$$

- yielding the a posteriori distribution of the parameters

$$\pi(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{\int p(\mathbf{y}|\mathbf{x}')p(\mathbf{x}')d\mathbf{x}'}$$

Hereafter

- the dependence upon the observations $\mathbf{y}$ is omitted: $\pi(\mathbf{x})$.
- the likelihood $p(\mathbf{y}|\mathbf{x})$ is normalized.
- it is assumed $\mathbf{x} \in \mathbb{R}^d$ and the prior has a density w.r.t. the Lebesgue measure.

Learn about the a posteriori distribution

- for parameter estimation: maximum a posteriori; mean a posteriori

$$\int \mathbf{x} \, \pi(\mathbf{x}) d\mathbf{x}.$$

- for model comparison

$$e(Y) = \int p(Y|\mathbf{x}) \; p(\mathbf{x}) d\mathbf{x} \qquad \text{evidence}$$

$$\frac{\int p_1(Y|\mathbf{x}) \; p_1(\mathbf{x}) d\mathbf{x}}{\int p_2(Y|\mathbf{x}) \; p_2(\mathbf{x}) d\mathbf{x}} \qquad \text{Bayes factor}$$

- for predictive inference

$$\int p(Y^\star|\mathbf{x}) \; \pi(\mathbf{x}) d\mathbf{x}.$$

$\hookrightarrow$ Interested in

- the exploration of the a posteriori distribution π
- the computation of integrals w.r.t. π

Unfeasibility

- dimension and complexity of the space:

  π is a distribution on $\mathsf{X} \subseteq \mathbb{R}^d$.
- π is (usually) known up to a normalizing constant

$$\pi(\mathbf{x}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{\int p(\mathbf{y}|\mathbf{x}')p(\mathbf{x}')d\mathbf{x}'} \propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$$

Unfeasibility

- dimension and complexity of the space:

  π is a distribution on X $\subseteq \mathbb{R}^d$.

- π is (usually) known up to a normalizing constant

$$\pi_u(\mathbf{x}) = \qquad\qquad\qquad p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$$

Unfeasibility

- dimension and complexity of the space:

  π is a distribution on X $\subseteq \mathbb{R}^d$.
- π is (usually) known up to a normalizing constant

$$\pi_u(\mathbf{x}) = \qquad\qquad p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$$

Therefore,

- exact exploration, exact integration are untractable.
- numerical approximation such as Monte Carlo methods is required.

Monte Carlo methods (1/2)

- Probabilistic approximation of a target distribution π - may be known up to a normalizing constant.

- Idea :
  - choose a proposal (trial, instrumental, $\cdots$) distribution, and draw at random points $X_1, \cdots, X_k, \cdots$
  - modify these points in order to obtain an approximation of π

Mecanism 1: associate a *weight* to each point.
Ex.: Importance Sampling

Mecanism 2: discard some points using an *acceptance-rejection* rule.
Ex.: Markov chain Monte Carlo

Monte Carlo methods (1/2)

- Probabilistic approximation of a target distribution π - may be known up to a normalizing constant.

- Idea :
  - choose a proposal (trial, instrumental, $\cdots$) distribution, and draw at random points $X_1, \cdots, X_k, \cdots$
  - modify these points in order to obtain an approximation of π

Mecanism 1: associate a *weight* to each point.
Ex.: Importance Sampling

Mecanism 2: discard some points using an *acceptance-rejection* rule.
Ex.: Markov chain Monte Carlo

- General and flexible algorithms. But the *convergence* and the *efficiency* of these methods depend upon the proposal distribution.

Monte Carlo methods (2/2)

- Convergence: when the number of draws tends to infinity, do the samples approximate the target π?

Monte Carlo methods (2/2)

- Convergence: when the number of draws tends to infinity, do the samples approximate the target π?

- Efficiency: control/quantify the approximation

Monte Carlo methods (2/2)

- Convergence: when the number of draws tends to infinity, do the samples approximate the target π?

- Efficiency: control/quantify the approximation

- Role of the proposal distribution in the efficiency of the algorithm.

Monte Carlo methods (2/2)

- Convergence: when the number of draws tends to infinity, do the samples approximate the target π?

- Efficiency: control/quantify the approximation

- Role of the proposal distribution in the efficiency of the algorithm.

- Adaptive methods for an automatic choice of the proposal distribution.

Outline

- Introduction
  - Bayesian Analysis
  - Monte Carlo methods
- Markov chain Monte Carlo
  - Algorithm
  - Convergence of the method
  - Accuracy of the approximation
  - Proposal distribution and efficiency
  - Adaptive MCMC
  - Interacting methods
- Importance Sampling
  - Algorithm
  - Convergence of the method
  - Monitoring the convergence
  - Adaptive Importance sampling
  - Curse of dimensionality
- Conclusion
  - MCMC vs Importance Sampling
  - Burn In
  - Parallelization
- Bibliography

Algorithm (Hastings-Metropolis) (1/2)

Let $q : \mathsf{X} \times \mathsf{X} \to \mathbb{R}^+$ be the density of a *transition kernel*

$$\int_A q(\mathbf{x},\mathbf{y})\, d\mathbf{y} = \text{probability of moving to the set } A\text{, starting from } \mathbf{x}$$

- Given the current sample X_k,
  1. draw a point $Y \sim q(X_k,\cdot)$
  2. accept or reject this point

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

where

$$\alpha(X_k,Y) = 1 \wedge \frac{\pi_u(Y)}{\pi_u(X_k)} \frac{q(Y,X_k)}{q(X_k,Y)}$$

Algorithm (Hastings-Metropolis) (1/2)

Let $q : \mathsf{X} \times \mathsf{X} \to \mathbb{R}^+$ be the density of a *transition kernel*

$$\int_A q(\mathbf{x},\mathbf{y})\, d\mathbf{y} = \text{probability of moving to the set } A\text{, starting from } \mathbf{x}$$

- Given the current sample X_k,
  1. draw a point $Y \sim q(X_k,\cdot)$
  2. accept or reject this point

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

where

$$\alpha(X_k,Y) = 1 \wedge \frac{\pi_u(Y)}{\pi_u(X_k)} \frac{q(Y,X_k)}{q(X_k,Y)}$$

- Approximate $\mathbb{E}_\pi\left[h(X)\right]$ by $\frac{1}{n}\sum_{k=1}^n h(X_k)$.

Algorithm (Hastings-Metropolis) (2/2)

▸ Biblio

- Independent HM: when q does not depend on the starting value $\mathbf{x}$. Then,

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

Algorithm (Hastings-Metropolis) (2/2)

▸ Biblio

- Independent HM: when q does not depend on the starting value $\mathbf{x}$. Then,

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

- Symmetric random walk HM: when q depends on $\mathbf{x}$,$\mathbf{y}$ through $\|\mathbf{x} - \mathbf{y}\|$. Then,

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

  - Ex. $q(\mathbf{x},\mathbf{y}) = \mathcal{N}_d(\mathbf{x},\Gamma)[\mathbf{y}]$
  - Proposed moved are on the form

$$Y = X_k + Z \qquad Z \sim q(z)$$

  - Any move to a point Y such that $\pi(Y) \geq \pi(X_k)$ is accepted.

Convergence of the method (1/2)

By construction, $(X_k)_k$ is a Markov chain. There exist results on

- Ergodicity for any $\mathbf{x}$,

$$\lim_{n\to\infty} \sup_{\{h:|h|\leq 1\}} \Big| \mathbb{E}\left[h(X_n)|X_0 = \mathbf{x}\right] - \mathbb{E}_\pi\left[h(X)\right] \Big| = 0.$$

Convergence of the method (1/2)

By construction, $(X_k)_k$ is a Markov chain. There exist results on

- V-Ergodicity for any $\mathbf{x}$,

$$\lim_{n\to\infty} \sup_{\{h:|h|\leq V\}} \Big| \mathbb{E}\left[h(X_n)|X_0=\mathbf{x}\right] - \mathbb{E}_\pi\left[h(X)\right] \Big| = 0.$$

- Explicit control of ergodicity

$$\sup_{\{h:|h|\leq V\}} \Big| \mathbb{E}\left[h(X_n)|X_0=\mathbf{x}\right] - \mathbb{E}_\pi\left[h(X)\right] \Big| \leq C\ r(n)\ V(x)$$

Convergence of the method (1/2)

By construction, $(X_k)_k$ is a Markov chain. There exist results on

- V-Ergodicity for any $\mathbf{x}$,

$$\lim_{n\to\infty} \sup_{\{h:|h|\leq V\}} \Big| \mathbb{E}\left[h(X_n)|X_0=\mathbf{x}\right] - \mathbb{E}_\pi\left[h(X)\right] \Big| = 0.$$

- Explicit control of ergodicity

$$\sup_{\{h:|h|\leq V\}} \Big| \mathbb{E}\left[h(X_n)|X_0=\mathbf{x}\right] - \mathbb{E}_\pi\left[h(X)\right] \Big| \leq C\ r(n)\ V(x)$$

- Law of large numbers

$$\lim_n \frac{1}{n}\sum_{k=1}^n h(X_k) = \mathbb{E}_\pi\left[h(X)\right] \qquad \text{a.s.}$$

- Central Limit Theorem, deviation inequalities, $\cdots$

$$\sqrt{n}\left|\frac{1}{n}\sum_{k=1}^n h(X_k) - \mathbb{E}_\pi\left[h(X)\right]\right| \xrightarrow{\mathcal{D}} \mathcal{N}(0,\Gamma)$$

Convergence of the method (2/2)

▸ Biblio

Such convergence results are established under assumptions

- on the target π and its support X
  - X is compact (simple theorey). Or not (a bit more technical!)
  - regularity on π
  - decaying rates of $\mathbf{x} \mapsto \pi(\mathbf{x})$ in the tails.
- on the function h (e.g. for CLT)

Convergence of the method (2/2)

▸ Biblio

Such convergence results are established under assumptions

- on the target π and its support X
  - X is compact (simple theorey). Or not (a bit more technical!)
  - regularity on π
  - decaying rates of $\mathbf{x} \mapsto \pi(\mathbf{x})$ in the tails.
- on the function h (e.g. for CLT)
- on the proposal kernel q
  - irreducibility of the chain
  - upper bounds and lower bounds

The user chooses the proposal - the convergence and the efficiency of the algorithm depends upon q.

Accuracy of the approximation: explicit control of ergodicity

▸ Biblio

$$\sup_{\{h:|h|\leq V\}} \Big| \mathbb{E}\left[h(X_n)|X_0 = \mathbf{x}\right] - \mathbb{E}_\pi\left[h(X)\right] \Big| \leq C\ r(n)\ V(\mathbf{x})$$

- Could be used to determine the length n of the chain to reach a fixed accuracy, depending upon the initial value $\mathbf{x}$.
- In practice, C is very large, $\lim_{|x|\to\infty} V(x) = +\infty \cdots$
- To my opinion, hopeless (given the current literature).

Accuracy of the approximation: variance in the CLT (1/2)

- When CLT holds, the limiting variance is

$$\begin{aligned}\sigma^2 &= \mathrm{Var}_\pi(h(X)) + 2\sum_{k\geq 1}\mathrm{Cov}_\pi\Big(h(X_0),h(X_k)\Big)\\ &= \gamma(0) + 2\sum_{k\geq 1}\underbrace{\gamma(k)}_{\text{lag } k \text{ autocovariance}} .\end{aligned}$$

- If $\lim_n \hat{\sigma}_n^2 = \sigma^2$ $_{\text{a.s. or } \mathbb{P}}$, we can form confidence interval with half size

$$t_\star \, \frac{\hat{\sigma}_n}{\sqrt{n}}, \qquad t_\star \text{ appropriate quantile}$$

↪ How to estimate σ^2 from the samples $X_1, \cdots, X_n$?

Accuracy of the approximation: variance in the CLT (2/2)

- Spectral methods

$$\hat{\sigma}_n^2 = \sum_{k=-b_n}^{b_n} \omega_n(k) \; \hat{\gamma}_n(k)$$

where

$$\hat{\gamma}_n(k) = \frac{1}{n} \sum_{\ell=1}^{n-|k|} \left(X_\ell - \frac{1}{n} \sum_{j=1}^{n} X_j \right) \left(X_{\ell+|k|} - \frac{1}{n} \sum_{j=1}^{n} X_j \right)$$

Is is a consistent estimator of σ^2 under conditions on

- the *lag window* $\omega_n(\cdot)$ and b_n. For example,
  - Truncation : $\omega_n(k) = 1$ if $|k| \leq b_n$ and 0 otherwise : NOT possible.
  - Parzen: $\omega_n(k) = 1 - |k|^q / b_n^q$ if $|k| \leq b_n$. $(q \in \mathbb{Z}_+)$.
  - Tukey-Hanning: $\omega_n(k) = 0.5(1 + \cos(\pi |k| / b_n))$ if $|k| \leq b_n$.
- the mixing properties of the chain uniform ergodicity, geometric ergodicity.

Accuracy of the approximation: variance in the CLT (2/2)

- Spectral methods
- (non overlapping) Batch means $n = a_n \, b_n$: a_n blocks of length b_n.

$$\hat{\sigma}_n^2 = \frac{b_n}{a_n - 1} \sum_{k=0}^{a_n - 1} \left(\underbrace{\frac{1}{b_n} \sum_{\ell=1}^{b_n} h(X_{k b_n + \ell})}_{\text{mean over block } k} - \underbrace{\frac{1}{n} \sum_{k=1}^{n} h(X_k)}_{\text{mean over the full path}} \right)^2$$

Is is a consistent estimator of σ^2 under conditions on

- the mixing properties of the chain
- a_n,b_n

  - both of them have to increase with n, at some rate.
  - this rate depends upon the mixing properties of the chain.

Accuracy of the approximation: variance in the CLT (2/2)

- Spectral methods
- (non overlapping) Batch means
- Overlapping batch means $n = (n - b_n + 1)$ overlapping batches of length b_n.

$$\hat{\sigma}_n^2 = \frac{n b_n}{(n - b_n + 1)(n - b_n)} \sum_{k=0}^{n-b_n} \left(\underbrace{\frac{1}{b_n} \sum_{\ell=1}^{b_n} h(X_{k+\ell})}_{\text{mean from } k \text{ to } k+b_n-1} - \underbrace{\frac{1}{n} \sum_{k=1}^{n} h(X_k)}_{\text{mean over the full path}} \right)^2$$

Is is a consistent estimator of σ^2 under conditions on

- the mixing properties of the chain
- a_n,b_n

  - both of them have to increase with n, at some rate.
  - this rate depends upon the mixing properties of the chain.

Accuracy of the approximation: variance in the CLT (2/2)

- Spectral methods
- (non overlapping) Batch means
- Overlapping batch means
- Regenerative simulation
  - Sample the chain in order to introduce some *regeneration times* $\tau_1, \cdots, \tau_{R_n}$
  - Estimate the variance by

$$\hat{\sigma}_n^2 = \frac{R_n}{\tau_{R_n}^2} \sum_{k=1}^{R_n} \left(\sum_{\ell=\tau_{k-1}+1}^{\tau_k} \{ h(X_\ell) - \left(\frac{1}{\tau_{R_n}} \sum_{j=1}^{\tau_{R_n}} h(X_j) \right) \} \right)^2$$

  - Consistency is established.
  - In practice, it is difficult to obtain many regeneration times.

Accuracy of the approximation: variance in the CLT (2/2)

- Spectral methods
- (non overlapping) Batch means
- Overlapping batch means
- Regenerative simulation

Based on empirical results,

$$\text{spectral, overlapping BM} > \text{BM} > \text{regenerative}$$

Note that these estimators can be used to stop a MCMC run with

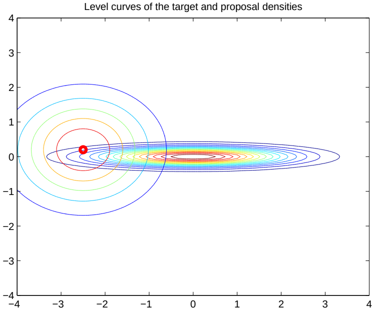
- a *fixed time rule*. Then, check if the confidence interval is undesirable wide or not.
- a *fixed width rule*: stop when the interval is sufficiently narrow.

▸ Biblio

Proposal distribution and efficiency (1/3)

- The direction of the moves

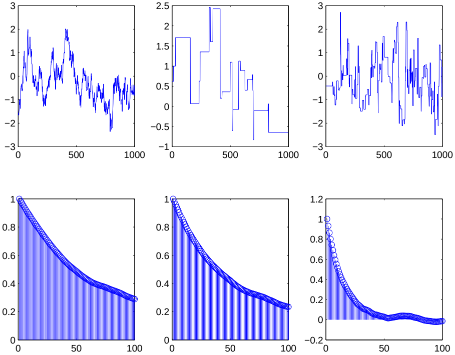
Symmetric Random Walk chain on $\mathbb{R}^2$, with target density $\mathcal{N}(0,\Gamma)$ and proposal distribution $\mathcal{N}(0,I)$



Proposal distribution and efficiency (2/3)

- The size of the moves

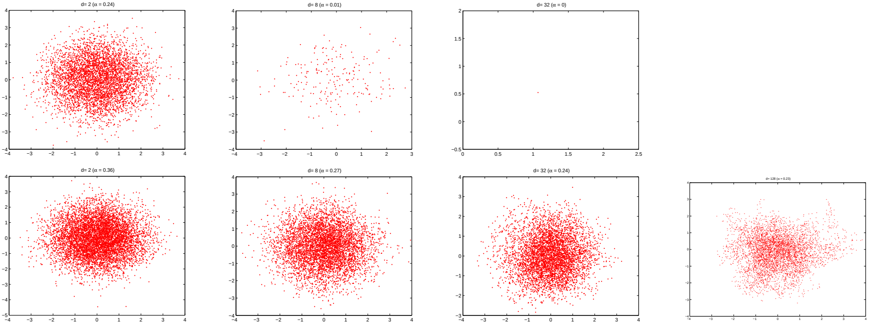
Symmetric Random Walk chain on $\mathbb{R}$, with Gaussian proposal of variance σ^2.



Three different values of σ : [top] a path of the chain [bottom] auto-correlation function

Proposal distribution and efficiency (3/3)

- The curse of dimensionality Symmetric Random Walk chain on $\mathbb{R}^d$, with target distribution $\mathcal{N}(0,I)$ and



$d \in \{2{,}8{,}32{,}64\}$: projection of the chain $(x_1, \cdots, x_d)$ on $\mathbb{R}^2$. [top] σ does not depend on d and $\overline{\alpha}$ is resp. 25%, 1%, 0. [bottom] σ is of the form $c/\sqrt{d}$ and $\overline{\alpha}$ is resp. 36%, 27%, 24% and 23%.

Optimal scaling

▸ Biblio

- Theoretical results:
  - study the skeleton process (when $d \to \infty$) associated to the chain.
  - optimize the speed of this process.

These results are obtained

- when the target π has independent marginals.
- when the chain is stationary : $X_0 \sim \pi$.

Optimal scaling

▸ Biblio

- Theoretical results:
  - study the skeleton process (when $d \to \infty$) associated to the chain.
  - optimize the speed of this process.

These results are obtained

- when the target π has independent marginals.
- when the chain is stationary : $X_0 \sim \pi$.

In the case of Sym. Random Walk HM with proposal $\mathcal{N}(0,c^2/d\ \Gamma)$

$$c_\star = 2.38^2 \qquad \Gamma_\star = \text{covariance matrix of the target } \pi$$

yielding to a so-called optimal mean acceptance-rejection ratio

$$\overline{\alpha}_\star = 0.234$$

- In practice:
  - What about the transient phase and "small" d?
  - The covariance matrix of π is unknown.

Adaptive MCMC (1/3)

▸ Biblio

- Pioneering work: *Adaptive Monte Carlo*
  - which is an adaptive Sym. Random Walk HM sampler,
  - start with an initial covariance matrix $\Gamma^{(0)}$ for the Gaussian proposal distribution,
  - update the covariance matrix $\Gamma^{(t)}$ at every iteration, or after a block of iterations, or $\cdots$ by using the past samples of the chain.
  - The convergence of this sampler is now established (LLN, CLT).

Adaptive MCMC (1/3)

▸ Biblio

▶ Pioneering work: *Adaptive Monte Carlo*

- which is an adaptive Sym. Random Walk HM sampler,
- start with an initial covariance matrix $\Gamma^{(0)}$ for the Gaussian proposal distribution,
- update the covariance matrix $\Gamma^{(t)}$ at every iteration, or after a block of iterations, or $\cdots$ by using the past samples of the chain.
- The convergence of this sampler is now established (LLN, CLT).

▶ Now, many adaptive MCMC algorithms for an automatic tuning of a *design parameter*

- define an accuracy criterion; usually no explicit optimum for this criterion
- update the parameter by using the current draws, in order to *asymptotically,* *when* $n \to \infty$, optimize this accuracy criterion.
  `tool for the update rule:` stochastic gradient algorithm, stochastic approximation alg., expectation-maximization alg., $\cdots$

Adaptive MCMC (2/3)

Unfortunately, adaptation can destroy the convergence to π !

Adaptive MCMC (2/3)

Unfortunately, adaptation can destroy the convergence to π !

- Let $\theta \in (0,1)$. Consider the transition matrix

$$P_\theta = \begin{pmatrix} 1-\theta & \theta \\ \theta & 1-\theta \end{pmatrix}$$

A Markov chain with this transition matrix converges to the stationary distribution $\pi = (1/2; 1/2)$.

Adaptive MCMC (2/3)

Unfortunately, adaptation can destroy the convergence to π !

- Let $\theta \in (0,1)$. Consider the transition matrix

$$P_\theta = \begin{pmatrix} 1-\theta & \theta \\ \theta & 1-\theta \end{pmatrix}$$

A Markov chain with this transition matrix converges to the stationary distribution $\pi = (1/2; 1/2)$.

- Fix $t_0,t_1 \in (0,1)$. Define a chain as follows: given X_k,

$$X_{k+1} \sim \begin{cases} P_{t_0}(X_k,\cdot) & \text{if } X_k = 0 \\ P_{t_1}(X_k,\cdot) & \text{if } X_k = 1 \end{cases}$$

Adaptive MCMC (2/3)

Unfortunately, adaptation can destroy the convergence to π !

- Let $\theta \in (0,1)$. Consider the transition matrix

$$P_\theta = \begin{pmatrix} 1-\theta & \theta \\ \theta & 1-\theta \end{pmatrix}$$

A Markov chain with this transition matrix converges to the stationary distribution $\pi = (1/2; 1/2)$.

- Fix $t_0, t_1 \in (0,1)$. Define a chain as follows: given X_k,

$$X_{k+1} \sim \begin{cases} P_{t_0}(X_k, \cdot) & \text{if } X_k = 0 \\ P_{t_1}(X_k, \cdot) & \text{if } X_k = 1 \end{cases}$$

- Then, $(X_n)_n$ is a Markov chain, with transition matrix

$$\begin{pmatrix} 1-t_0 & t_0 \\ t_1 & 1-t_1 \end{pmatrix}$$

but it converges to the distribution $\tilde{\pi} \propto (t_1, t_0) \neq \pi$.

Adaptive MCMC (3/3)

▸ Biblio

- In Adaptive MCMC, there is a family of kernel $(P_\theta, \theta \in \Theta)$ and all these kernels have the same invariant distribution π.
- At each iteration, pick one of this kernel P_{θ_k} with a random mecanism e.g. depending upon the past samples.
- The resulting chain is not necessarily a Markov chain, and may converge to a distribution $\tilde{\pi} \neq \pi$.

Adaptive MCMC (3/3)

▸ Biblio

- In Adaptive MCMC, there is a family of kernel $(P_\theta, \theta \in \Theta)$ and all these kernels have the same invariant distribution π.
- At each iteration, pick one of this kernel P_{θ_k} with a random mecanism e.g. depending upon the past samples.
- The resulting chain is not necessarily a Markov chain, and may converge to a distribution $\tilde{\pi} \neq \pi$.

▶ Sufficient conditions for the convergence (convergence to π, Law of large numbers, CLT) of adaptive algorithms. Essentially,

- **Diminishing adaption:** $d(P_{\theta_k}, P_{\theta_{k+1}}) \to 0$ at some rate, in some sense.
- **Containment condition:** the transition kernels $(P_\theta, \theta \in \Theta)$ have a *similar* ergodic behavior.

Interacting methods (1/2)

▸ Biblio

Due to

- the curse of dimensionality
- the multimodality of the target π

new MCMC methodologies are about interacting algorithms

▶ Idea:

- Run K chains in parallel, each with its own invariant distribution $\pi^{(k)}$ by allowing interaction between neighboring chains.
- $\pi^{(k)}$ chosen so that the associated chain has good mixing properties. And $\pi^{(K)} = \pi$.

Interacting methods (1/2)

▸ Biblio

Due to

- the curse of dimensionality
- the multimodality of the target π

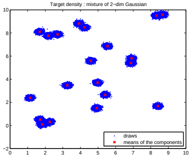
new MCMC methodologies are about interacting algorithms

▶ Idea:

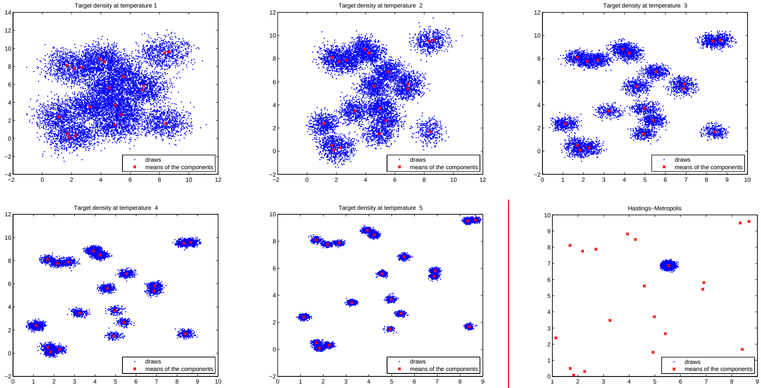
- Run K chains in parallel, each with its own invariant distribution $\pi^{(k)}$ by allowing interaction between neighboring chains.
- $\pi^{(k)}$ chosen so that the associated chain has good mixing properties. And $\pi^{(K)} = \pi$.
- Ex. $\pi^{(k)}$ is a tempered version of π. `Tempering, Equi-Energy sampler, Wang-Landau,` $\cdots$ `many ideas from numerical Statistical Physics and Molecular Dynamics`

▶ Convergence results: Few answers, mainly an open question !

Interacting methods (2/2)



- Target: $\pi = \sum_{i=1}^{20} \mathcal{N}_2(\mu_i, \Sigma_i)$
- $\pi^{(k)} = \pi^{1/T_k}$ $\quad T_1 > T_2 > \cdots > T_K = 1$



Outline

- Introduction
  - Bayesian Analysis
  - Monte Carlo methods
- Markov chain Monte Carlo
  - Algorithm
  - Convergence of the method
  - Accuracy of the approximation
  - Proposal distribution and efficiency
  - Adaptive MCMC
  - Interacting methods
- Importance Sampling
  - Algorithm
  - Convergence of the method
  - Monitoring the convergence
  - Adaptive Importance sampling
  - Curse of dimensionality
- Conclusion
  - MCMC vs Importance Sampling
  - Burn In
  - Parallelization
- Bibliography

Algorithm

Choose a proposal distribution $q(\mathbf{x})$.

1. Draw independently points $X_1, \cdots, X_n, \cdots$ under q.
2. Compute an importance weight for each point
$$\omega_k = \frac{\pi_u(X_k)}{q(X_k)}$$
3. Approximate π by the weighted points
$$\int h(\mathbf{x})\ \pi(\mathbf{x}) d\mathbf{x} = \mathbb{E}_\pi\left[h(X)\right] \approx \sum_{k=1}^{n} \frac{\omega_k}{\sum_{\ell=1}^{n} \omega_\ell}\ h(X_k)$$

When the normalizing constant of π is known, replace this approximation with

$$\frac{1}{n} \sum_{k=1}^{n} \omega_k\ h(X_k).$$

Hereafter, only the case "π is known up to a normalizing constant" is considered

Convergence of the method (1/4)

- Consistent estimator

For any function h s.t. $\mathrm{Supp}(\pi|h|) \subset \mathrm{Supp}(q)$ *

$$\lim_{n\to\infty} \sum_{k=1}^{n} \frac{\omega_k}{\sum_{\ell=1}^{n} \omega_\ell} h(X_k) \xrightarrow{a.s.} \int h(\mathbf{x})\ \pi(\mathbf{x}) d\mathbf{x}$$

which implies that

$$\int_\Delta \pi(\mathbf{x}) d\mathbf{x} \approx \sum_{k=1}^{n} \frac{\omega_k}{\sum_{\ell=1}^{n} \omega_\ell}\ \mathbb{1}_\Delta(X_k)$$

* for example, choose q so that $\{q=0\} \subseteq \{\pi|h|=0\}$

Convergence of the method (2/4)

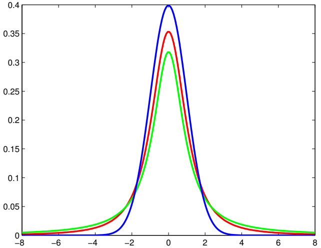
- Toy example

$$\text{compute} \int_{\mathbb{R}} |x|\pi(x)dx \qquad \text{when} \qquad \pi(x) \sim t(3) \propto \frac{1}{(1+\frac{x^2}{3})^2}$$

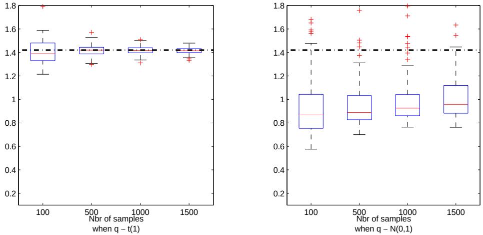
Consider in turn the proposal q equal to

a Student $t(1)$

a Normal $\mathcal{N}(0{,}1)$



Plot of the densities q (green, blue) and π (in red)



Boxplot computed from 100 runs of the algorithm

Convergence of the method (2/4)

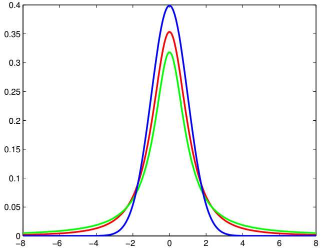
- Toy example

$$\text{compute} \int_{\mathbb{R}} |x| \pi(x) dx \qquad \text{when} \qquad \pi(x) \sim t(3) \propto \frac{1}{(1+\frac{x^2}{3})^2}$$

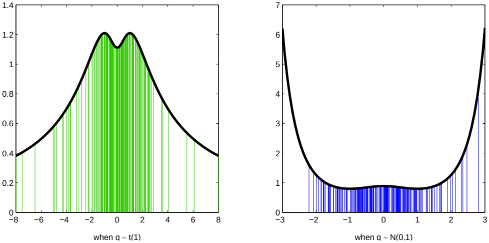
Consider in turn the proposal q equal to

a Student $t(1)$

a Normal $\mathcal{N}(0,1)$



Plot of the densities q (green, blue) and π (in red)



On one run of the algorithm :

weights of the draws (blue) and $x \mapsto \frac{\pi(x)}{q(x)}$ (black)

Convergence of the method (2/4)

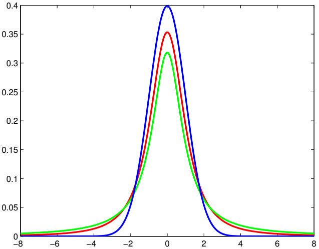
- Toy example

$$\text{compute} \int_{\mathbb{R}} |x|\pi(x)dx \qquad \text{when} \qquad \pi(x) \sim t(3) \propto \frac{1}{(1+\frac{x^2}{3})^2}$$

Consider in turn the proposal q equal to

a Student $t(1)$

a Normal $\mathcal{N}(0{,}1)$



The efficiency of the algorithm depends upon the proposal distribution q: if few large weights and the others negligible, the approximation is likely not accurate

Plot of the densities q (green, blue) and π (in red)

Convergence of the method (3/4)

- Variance of the estimator

$$\mathrm{Var}\left(\sum_{k=1}^{n} \frac{\omega_k}{\sum_{\ell=1}^{n} \omega_\ell} h(X_k)\right) = n^{-1}\,\sigma^2 + o\left(\frac{1}{n}\right)$$

with

$$\sigma^2 = \mathbb{E}_\pi\left[(h(\mathrm{X}) - \mathbb{E}_\pi\left[h(X)\right])^2\,\frac{\pi(\mathrm{X})}{q(\mathrm{X})}\right]$$

Note that, as a function of q, σ^2 is minimal by choosing q as a function of π, h namely

$$q_\star \propto \left|h - \mathbb{E}_\pi\left[h(X)\right]\right|\,\pi$$

Rule of thumb: choose the proposal so that

$$\sup_{\mathbf{x}} \frac{\pi_u(\mathbf{x})}{q(\mathbf{x})} < \infty.$$

q has heavier tails than π;

q does not depend on h.

Convergence of the method (4/4)

- Asymptotic normality

$$\sigma^2 = \mathbb{E}_\pi \left[(h(\mathrm{X}) - \mathbb{E}_\pi \left[h(X) \right])^2 \, \frac{\pi(\mathrm{X})}{q(\mathrm{X})} \right]$$

It holds:

$$\lim_n n \sum_{k=1}^{n} \left(\frac{\omega_k}{\sum_{\ell=1}^{n} \omega_\ell} \right)^2 \left(h(X_k) - \sum_{j=1}^{n} \frac{\omega_j}{\sum_{\ell=1}^{n} \omega_\ell} h(X_j) \right)^2 = \sigma^2,$$

so that

- it is possible to estimate the asymptotic variance from the samples.

Convergence of the method (4/4)

- Asymptotic normality

$$\sigma^2 = \mathbb{E}_\pi \left[(h(\mathrm{X}) - \mathbb{E}_\pi \left[h(X)\right])^2 \frac{\pi(\mathrm{X})}{q(\mathrm{X})} \right]$$

It holds:

$$\lim_n n \sum_{k=1}^n \left(\frac{\omega_k}{\sum_{\ell=1}^n \omega_\ell} \right)^2 \left(h(X_k) - \sum_{j=1}^n \frac{\omega_j}{\sum_{\ell=1}^n \omega_\ell} h(X_j) \right)^2 = \sigma^2,$$

and

$$\sqrt{n} \left(\sum_{k=1}^n \frac{\omega_k}{\sum_{\ell=1}^n \omega_\ell} h(X_k) - \mathbb{E}_\pi \left[h(X)\right] \right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0,\sigma^2\right)$$

so that

- it is possible to estimate the asymptotic variance from the samples.

- (asymptotic) confidence intervals for the approximation of $\mathbb{E}_\pi[h(X)]$.

Monitoring the convergence: Coefficient of Variation

$$\mathrm{CV}_n = \sqrt{n \sum_{k=1}^{n} \left(\frac{\omega_k}{\sum_{\ell=1}^{n} \omega_\ell} - \frac{1}{n} \right)^2}$$

- a measure of the number of ineffective particles:
  - CV_n is minimal $(= 0)$ when the weights are equal.
  - CV_n is maximal $(= \sqrt{n-1})$ when all weights are null but one.

Monitoring the convergence: Coefficient of Variation

$$\mathrm{CV}_n = \sqrt{n \sum_{k=1}^{n} \left(\frac{\omega_k}{\sum_{\ell=1}^{n} \omega_\ell} - \frac{1}{n} \right)^2}$$

- a measure of the number of ineffective particles:
    - CV_n is minimal $(= 0)$ when the weights are equal.
    - CV_n is maximal $(= \sqrt{n-1})$ when all weights are null but one.

- When $n \to \infty$,

$$\lim_n \mathrm{CV}_n = D_{\chi^2}(\pi, q) \qquad (\textit{Pearson-}\chi^2 \textit{ distance})$$

where

$$\left(D_{\chi^2}(\pi,q)\right)^2 = \int \left(\frac{\pi(\mathrm{x})}{q(\mathrm{x})} - 1 \right) \pi(\mathrm{x}) d\mathrm{x} = \mathrm{Var}_q \left(\frac{\pi(X)}{q(X)} \right).$$

Monitoring the convergence: Effective Sample Size

$$\mathrm{ESS}_n = \left(\sum_{k=1}^{n} \left(\frac{\omega_k}{\sum_{\ell=1}^{n} \omega_\ell} \right)^2 \right)^{-1} = \frac{n}{1 + \mathrm{CV}_n}$$

- a measure of the number of effective particles:
  - ESS_n is maximal $(= n)$ when the weights are equal.
  - ESS_n is minimal $(= 1)$ when all weights are null but one.

Monitoring the convergence: Effective Sample Size

$$\mathrm{ESS}_n = \left(\sum_{k=1}^{n} \left(\frac{\omega_k}{\sum_{\ell=1}^{n} \omega_\ell} \right)^2 \right)^{-1} = \frac{n}{1 + \mathrm{CV}_n}$$

- a measure of the number of effective particles:
  - ESS_n is maximal $(= n)$ when the weights are equal.
  - ESS_n is minimal $(= 1)$ when all weights are null but one.

- Heuristically,

$$\frac{\mathrm{Var}_\pi(h)}{\sigma^2} \approx \frac{1}{1 + \mathrm{Var}_q\left(\frac{\pi(X)}{q(X)}\right)} = \lim_n \frac{1}{1 + \mathrm{CV}_n},$$

Asymptotically, the number of points of i.i.d. samples drawn from π equivalent to the n weighted samples in terms of accuracy is

$$n \frac{\mathrm{Var}_\pi(h)}{\sigma^2}$$

Monitoring the convergence: Effective Sample Size

$$\mathrm{ESS}_n = \left(\sum_{k=1}^{n} \left(\frac{\omega_k}{\sum_{\ell=1}^{n} \omega_\ell} \right)^2 \right)^{-1} = \frac{n}{1 + \mathrm{CV}_n}$$

- a measure of the number of effective particles:
  - ESS_n is maximal $(= n)$ when the weights are equal.
  - ESS_n is minimal $(= 1)$ when all weights are null but one.

- Heuristically,

$$\frac{\mathrm{Var}_\pi(h)}{\sigma^2} \approx \frac{1}{1 + \mathrm{Var}_q\left(\frac{\pi(X)}{q(X)}\right)} = \lim_n \frac{1}{1 + \mathrm{CV}_n},$$

Asymptotically, the number of points of i.i.d. samples drawn from π equivalent to the n weighted samples to achieve a fixed accuracy

$$n \frac{\mathrm{Var}_\pi(h)}{\sigma^2} = \mathrm{ESS}_n$$

Monitoring the convergence: Normalized perplexity

$$\mathcal{E}_n = \frac{1}{n} \exp\left(-\sum_{i=1}^{n} \frac{\omega_i}{\sum_{\ell=1}^{n} \omega_\ell} \, \log\left(\frac{\omega_i}{\sum_{\ell=1}^{n} \omega_\ell} \right) \right)$$

- The normalized perplexity is
  - maximal $(= 1)$ when the weights are equal.
  - minimal $(= 1/n)$ when all weights are zero but one.

Monitoring the convergence: Normalized perplexity

$$\mathcal{E}_n = \frac{1}{n} \exp\left(- \sum_{i=1}^{n} \frac{\omega_i}{\sum_{\ell=1}^{n} \omega_\ell} \; \log\left(\frac{\omega_i}{\sum_{\ell=1}^{n} \omega_\ell} \right) \right)$$

- The normalized perplexity is
  - maximal $(= 1)$ when the weights are equal.
  - minimal $(= 1/n)$ when all weights are zero but one.

- As $n \to +\infty$,

$$\begin{aligned} \lim_n \mathcal{E}_n &= \exp\left(- \int \log\left(\frac{\pi(\mathbf{x})}{q(\mathbf{x})} \right) \; \pi(\mathbf{x}) \; d\mathbf{x} \right) \\ &= \exp\left(-d_{\mathrm{KL}} \left(\pi , q \right) \right) \qquad \text{(Kullback-Leibler divergence)} \end{aligned}$$

$\mathcal{E}_n$ is a measure of fit of the proposal distribution q.

Adaptive Importance sampling (1/3)

▸ Biblio

- The choice of q is crucial for the efficiency of Importance Sampling.
- Methods were proposed to reach the objective:

  choose the distribution q in a family of densities $\mathcal{Q}$, as the optimum of an adequacy criterion

Adaptive Importance sampling (1/3)

▸ Biblio

- The choice of q is crucial for the efficiency of Importance Sampling.
- Methods were proposed to reach the objective:

  choose the distribution q in a family of densities $\mathcal{Q}$, as the optimum of an adequacy criterion

▶ Example (Population Monte Carlo): solve

$$\mathrm{argmin}_{q \in \mathcal{Q}} \, d_{\mathrm{KL}}(\pi, q) = \mathrm{argmin}_{q \in \mathcal{Q}} \int \log \frac{\pi(\mathbf{x})}{q(\mathbf{x})} \, \pi(\mathbf{x}) d\mathbf{x}$$

Adaptive Importance sampling (1/3)

▸ Biblio

- The choice of q is crucial for the efficiency of Importance Sampling.
- Methods were proposed to reach the objective:

  choose the distribution q in a family of densities $\mathcal{Q}$, as the optimum of an adequacy criterion

▶ Example (Population Monte Carlo): solve

$$\mathrm{argmin}_{q \in \mathcal{Q}} \, d_{\mathrm{KL}} \left(\pi , q\right) = \mathrm{argmin}_{q \in \mathcal{Q}} \int \log \frac{\pi(\mathbf{x})}{q(\mathbf{x})} \; \pi(\mathbf{x}) d\mathbf{x}$$

▶ Example (Cross-Entropy method): solve

$$\mathrm{argmin}_{q \in \mathcal{Q}} \;\; d_{\mathrm{KL}} \left(\frac{|h| \pi}{\int |h(\mathbf{x})| \pi(\mathbf{x}) d(\mathbf{x})} , q \right)$$

Adaptive Importance sampling (1/3)

▸ Biblio

- The choice of q is crucial for the efficiency of Importance Sampling.
- Methods were proposed to reach the objective:

  choose the distribution q in a family of densities $\mathcal{Q}$, as the optimum of an adequacy criterion

▶ Example (Population Monte Carlo): solve

$$\mathrm{argmin}_{q \in \mathcal{Q}} \, d_{\mathrm{KL}} \left(\pi, q\right) = \mathrm{argmin}_{q \in \mathcal{Q}} \int \log \frac{\pi(\mathbf{x})}{q(\mathbf{x})} \; \pi(\mathbf{x}) d\mathbf{x}$$

▶ Example (Cross-Entropy method): solve

$$\mathrm{argmin}_{q \in \mathcal{Q}} \;\; d_{\mathrm{KL}} \left(\frac{|h| \pi}{\int |h(\mathbf{x})| \pi(\mathbf{x}) d(\mathbf{x})}, q \right)$$

- Nevertheless, (most of) the adequacy criterions depends on integrals w.r.t. π, which is precisely what we are not able to compute.

Adaptive Importance sampling (2/3)

Therefore, determine the *optimal* proposal distribution q adaptively:
▶ Example (Population Monte Carlo) - to follow

$$\mathrm{argmin}_{q \in \mathcal{Q}} \int \log \frac{\pi(\mathbf{x})}{q(\mathbf{x})} \, \pi(\mathbf{x}) d\mathbf{x} \Longleftrightarrow \mathrm{argmax}_{q \in \mathcal{Q}} \int \log q(\mathbf{x}) \, \pi(\mathbf{x}) d\mathbf{x}$$

1. Choose an initial distribution $q^{(0)}$, and compute an Importance Sampling approximation of the criterion

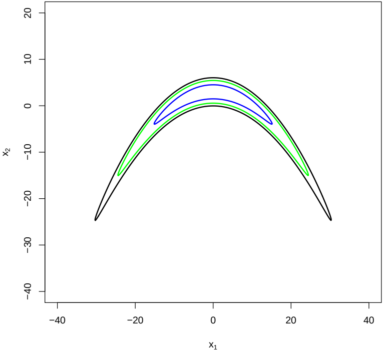
$$\sum_{k=1}^{n} \omega_k^{(0)} \log q(X_k)$$

2. Udpate the proposal: $q^{(1)}$ is an optimum of the approximated criterion.
3. Repeat until convergence.

In this example, Step 2 is explicit when $\mathcal{Q}$ is the family of mixture of Gaussian distributions, or mixture of t-distributions.

Adaptive Importance sampling (3/3)

- Population Monte Carlo - numerical application The target distribution in $\mathbb{R}^{10}$. Below marginal distribution of (x_1,x_2)



and $(x_3,\cdots,x_{10})$ are independent $\mathcal{N}(0,1)$.

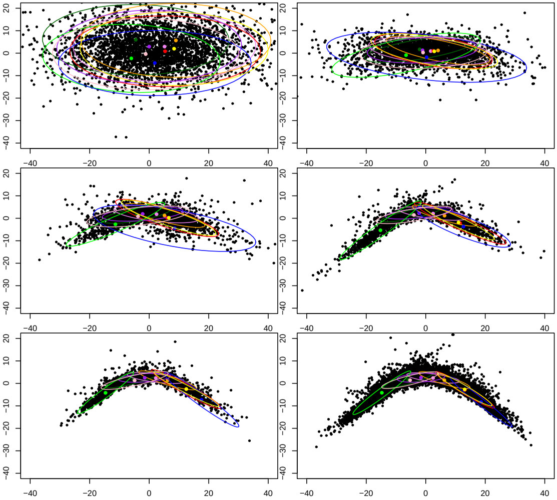


FIG.: Iterations 1,3,5,7,9,11. 10k points per plot, except 100k in the lase one. Mixture of 9 t-distributions, with 9 degrees of freedom

Monitoring convergence: the *Normalized perplexity (top panel)* and the *Normalized Effective Sample size* (bottom panel)

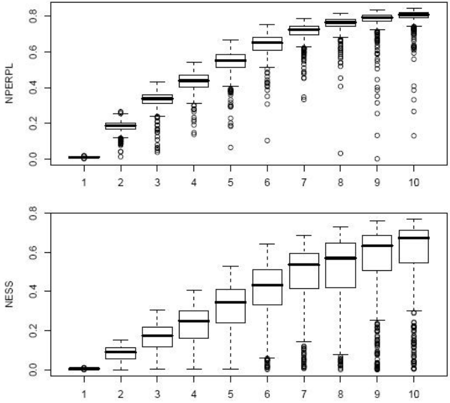
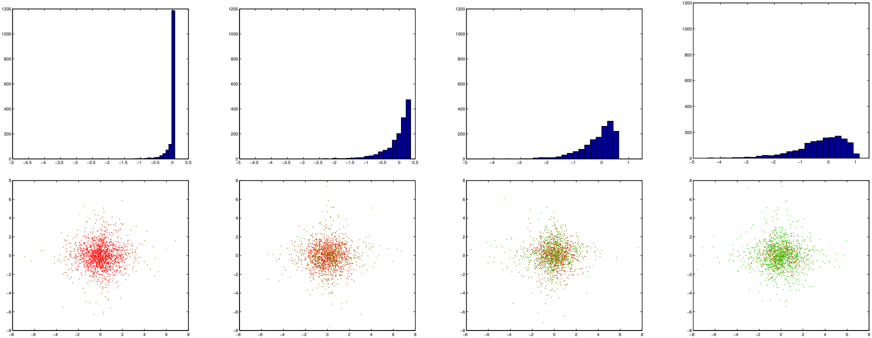


FIG.: for the first 10 iterations, over 500 simulation runs.

Curse of dimensionality

Is Importance Sampling robust to the dimension of the sampling space?

$$\pi(x_1,\cdots,x_d)=\prod_{k=1}^{d} t_4(x_k) \qquad q(x_1,\cdots,x_d)=\prod_{k=1}^{d} t_2(x_k)$$



(left to right) $d = 2{,}10{,}20{,}40$.
(top) Histogram of the log-weights (bottom) Draws - in the $(x_1{,}x2)$ plane; the color is prop. to the weight.

Introduction
- Bayesian Analysis
- Monte Carlo methods

Markov chain Monte Carlo
- Algorithm
- Convergence of the method
- Accuracy of the approximation
- Proposal distribution and efficiency
- Adaptive MCMC
- Interacting methods

Importance Sampling
- Algorithm
- Convergence of the method
- Monitoring the convergence
- Adaptive Importance sampling
- Curse of dimensionality

Conclusion
- MCMC vs Importance Sampling
- Burn In
- Parallelization

Bibliography

MCMC vs Importance Sampling

- Computational cost: (e.g. for the evaluation of π)

  - MCMC can not be parallelized , well, most of them
  - Importance Sampling allows for parallel computation.

- Monitoring the convergence

  - Importance Sampling: simple tools (CV, ESS, Perplexity)
  - MCMC: estimators of the asymptotic variance

- Proposal distribution

  Both the methods depend upon this design parameter $\longrightarrow$ adaptive algorithms.

- Curse of dimensionality

  MCMC more robust than Importance Sampling.

Burn In in MCMC

▸ Biblio

- The chain is started at X_0 which is not drawn under π.
- Hence, there is a bias:

$$\mathbb{E}\left[h(X_k)\right] \neq \mathbb{E}_\pi\left[h(X)\right], \qquad \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} h(X_k)\right] \neq \mathbb{E}_\pi\left[h(X)\right].$$

and discarding the first sample $X_1, \cdots, X_B$ can reduce the bias.

Burn In in MCMC

▸ Biblio

- The chain is started at X_0 which is not drawn under π.
- Hence, there is a bias:

$$\mathbb{E}\left[h(X_k)\right] \neq \mathbb{E}_\pi\left[h(X)\right], \qquad \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} h(X_k)\right] \neq \mathbb{E}_\pi\left[h(X)\right].$$

  and discarding the first sample $X_1, \cdots, X_B$ can reduce the bias.
- But it is possible (even likely) that

$$\mathrm{Var}\left(\frac{1}{n-B}\sum_{k=B}^{n} h(X_k)\right) \geq \mathrm{Var}\left(\frac{1}{n}\sum_{k=1}^{n} h(X_k)\right);$$

  the variance increases for the same computational cost n

Burn In in MCMC

▸ Biblio

- The chain is started at X_0 which is not drawn under π.
- Hence, there is a bias:

$$\mathbb{E}\left[h(X_k)\right] \neq \mathbb{E}_\pi\left[h(X)\right], \qquad \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} h(X_k)\right] \neq \mathbb{E}_\pi\left[h(X)\right].$$

  and discarding the first sample $X_1, \cdots, X_B$ can reduce the bias.

- But it is possible (even likely) that

$$\mathrm{Var}\left(\frac{1}{n-B}\sum_{k=B}^{n} h(X_k)\right) \geq \mathrm{Var}\left(\frac{1}{n}\sum_{k=1}^{n} h(X_k)\right);$$

  the variance increases for the same computational cost n

- Trade off ... Open question !

Parallelization

▸ Biblio

- Importance Sampling
  - YES ! sampling and computing the importance weights can easily be parallelized.
- MCMC
  - Part of independent-HM can be parallelized. Otherwise, difficult due to the Markov chain structure of the process.
  - One long run or r parallel chains?
    - there is values in trying a variety of initial distributions. E.g.: for multimodal target, with r starting points widely dispersed, better chance to recover the modes.
    - Parallel chains are superior if initialized from a distribution close to π.
    - r has to be large for an efficient estimation of the variance.
    - for a fixed computational cost N and with the same burn in B: $N - B$ points vs r chains with $(N - B)/r$ points.

Open question !

Books or survey on MCMC and Importance Sampling

Marin, J.M. and Robert, C. (2007). Bayesian Core: A practical approach to computational Bayesian analysis. Springer-Verlag, New York.

Robert, C. and Casella, G. (2004). Monte Carlo Statistical Methods, 2nd ed. Springer, New York.

Liu, J.S. (2008). Monte Carlo strategies in Scientific computing. Springer.

Roberts, G. and Rosenthal, J. (2004). General state space Markov chains and MCMC algorithms. Probab. Surv. 1:20-71

Rubinstein, R. and Kroese, D. (2008). Simulation and the Monte Carlo method. 2nd ed. Wiley Series in Probability and Statistics, Wiley-Interscience, Hoboken, NJ.

On the theory of Markov chains

Meyn, S. and Tweedie, R. (1993). Markov Chains and Stochastic Stability. Springer, London.

R. Douc, G. Fort, E. Moulines and P. Soulier (2004). Practical drift conditions for subgeometric rates of convergence. Ann. Appl. Probab. 14:1353-1377.

Jarner, S.F. and Roberts, G.O. (2002). Polynomial convergence rates of Markov chains. Ann. Appl. Probab. 12:224:247

Jones, G.L. (2004) On the Markov chain Central Limit Theorem. Probab. Surv. 1:299-320.

Nummelin, E. (1984). General Irreducible Markov Chains and Non-Negative Operators. Cambridge Univ. Press.

Roberts, G. and Rosenthal, J.S. (2004). General State space Markov chains and MCMC algorithms. Probab. Surv. 1:20-71.

On the convergence of MCMC samplers (Hastings-Metropolis)

Fort, G. and Moulines, E. (2000). V-subgeometric ergodicity for a Hastings-Metropolis algorithm. Statist. Probab. Lett. 49:401-410.

Fort, G. and Moulines, E. and Roberts, G.O. and Rosenthal, J.S. (2003) On the geometric ergodicity of hybrid samplers. J. Appl. Probab. 40:123-146.

Geyer, C. J. (1999). Likelihood inference for spatial point processes. In Stochastic Geometry: Likelihood and Computation (O. E. Barndorff-Nielsen, W. S. Kendall and M. N. M. van Lieshout, eds.) 79-140. Chapman & Hall/CRC, Boca Raton.

Jarner, S. and Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms. Stochastic Process. Appl. 85:341-361.

Mengersen, K. and Tweedie, R. (1996). Rates of convergence of the Hastings and Metropolis algorithms. Ann. Statist. 24:101-121.

Neath, R. and Jones, G. L. (2009). Variable-at-a-time implementations of Metropolis-Hastings. Technical report, School of Statistics, Univ. Minnesota.

Roberts, G. and Tweedie, R. (1996). Geometric convergence and central limit theorem for multidimensional Hastings and Metropolis algorithms. Biometrika 83:95-110.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). Ann. Statist. 22:1701-1762.

On the convergence of MCMC samplers (Gibbs)

Hobert, J. P. and Geyer, C. J. (1998). Geometric ergodicity of Gibbs and block Gibbs samplers for a hierarchical random effects model. J. Multivariate Anal. 67:414-430.

Hobert, J. P., Jones, G. L., Presnell, B. and Rosenthal, J. S. (2002). On the applicability of regenerative simulation in Markov chain Monte Carlo. Biometrika 89:731-743.

Jones, G. L. and Hobert, J. P. (2004). Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. Ann. Statist. 32:784-817.

Roberts, G. O. and Polson, N. G. (1994). On the geometric convergence of the Gibbs sampler. J. Roy. Statist. Soc. Ser. B 56:377-384.

Roberts, G. O. and Rosenthal, J. S. (1999). Convergence of slice sampler Markov chains. J. R. Stat. Soc. Ser. B Stat. Methodol. 61:643-660.

Rosenthal, J. S. (1996). Analysis of the Gibbs sampler for a model related to James-Stein estimators. Stat. Comput. 6:269-275.

Tan, A. and Hobert, J. P. (2009). Block Gibbs sampling for Bayesian random effects models with improper priors: Convergence and regeneration. J. Comput. Graph. Statist. 18:861-878.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). Ann. Statist. 22:1701-1762.

Explicit control of ergodicity

Baxendale, P. (2005) Renewal theory and computable convergence rates for geometrically ergodic Markov chains. Ann. Appl. Probab., 15:700-738.

Douc, R. and Moulines, E. and Rosenthal, J.S. (2004) Quantitative bounds for geometric convergence rates of Markov chains. Ann. Appl. Probab., 14:1643-1665.

G. Fort. (2002) Computable bounds for V-geometric ergodicity of Markov transition kernels. Technical Report RR 1047-M, Univ. J. Fourier, France.

Kolassa, J.E. (2000) Explicit bounds for geometric convergence of Markov chains. J. Appl. Probab., 37-642:651.

Mengersen, K.L. and Tweedie, R.L. (1996) Rates of convergence of the Hastings and Metropolis algorithms. Ann. Statist., 24:101-121.

Meyn, S.P. and Tweedie, R.L. (1994) Computable bounds for geometric convergence rates of Markov chains. Ann. Appl. Probab., 4:981-1011.

Roberts, G.O. and Tweedie, R.L. . (1999) Bounds on regeneration times and convergence rates for Markov Chains. Stochastic Process. Appl., 80:211-229.

Rosenthal, J.S. (1995) Minorization conditions and convergence rates for Markov chain Monte Carlo. J. Amer. Statist. Assoc., 90:558-566.

Estimating the variance in the CLT

Bratley, P. and Fox, B.L. and Schrage, L.E. (1987). A guide to simulation. Springer, New-York.

Flegal, J.M. and Jones, G.L. (2010). Batch Means and Spectra Variance Estimators in Markov Chain Monte Carlo. Ann. Statist. 38:1034-1070.

Hobert, J.P. and Jones, G.L. and Presnell, B. and Rosenthal, J.S. (2002) On the applicability of regenerative simulation in MArkov chain Monte Carlo. Biometrika 89:731-743.

Jones, G.L. and Haran, M. and Caffo, B.S. and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. J. Amer. Statist. Assoc. 101:1537-1547.

Jones, G.L. and Hobert, J.P. (2001) Honest exploration in intractable probability distributions via Markov chain Monte Carlo. Statist. Sci. 16:312-334.

Scaling

Bédard, M. (2006). Optimal acceptance rates for Metropolis algorithms: Moving beyond 0.234. Stochastic Process. Appl. 118:2198-222.

Breyer, L.A. and Roberts, G.O. (2000). From Metropolis to diffusions: Gibbs states and optimal scaling. Stochastic Process. Appl. 90:181-206.

Christensen, O.F. and Roberts, G. O. and Rosenthal, J.S. (2003). Scaling limits for the transient phase of local Metropolis-Hastings algorithms. J. R. Stat. Soc. Ser. B Stat. Methodol. 67:253-269.

Gelman, A. and Roberts, G.O. and Gilks, W.R. (1996). Efficient Metropolis jumping rules. Bayesian Statistics V, 599-608, Clarendon Press, Oxford.

Roberts, G. O. and Gelman, A. and Gilks, W.R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann. Appl. Probab. 7:110-20.

Roberts, G.O. and Rosenthal, J.S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. J. R. Stat. Soc. Ser. B Stat. Methodol. 60:255-268.

Roberts, G.O. and Rosenthal, J.S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. Statist. Sci. 16:351-367.

Adaptive MCMC algorithms (survey)

Andrieu, C. and Robert, C. (2001). Controlled markov chain monte carlo methods for optimal sampling. Tech. Rep. 125, Cahiers du Ceremade.

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. Statistics and Computing 18:343-373.

Atchade, Y. and Fort, G. and Moulines, E. and Priouret, P. (2011) Adaptive Markov chain Monte Carlo: Theory and Methods. Bayesian Time Series Models, Cambridge Univ. Press, Chapter 2, 33-53.

Atchade, Y.F. and Rosenthal, J.S. (2005). On adaptive Markov chain Monte Carlo algorithm. Bernoulli 11:815-828.

Roberts, G. and Rosenthal, J. (2009). Examples of adaptive MCMC. J. Comp. Graph. Stat. 18:349-367.

Rosenthal, J. S. (2009). MCMC Handbook, chap. Optimal Proposal Distributions and Adaptive MCMC. Chapman & Hall/CRC Press.

Adaptive MCMC algorithms (specific algorithms)

Bai, Y. and Craiu, R.V. and Di Narzo, A.F. (2010) Divide and Conquer: A Mixture-Based Approach to Regional Adaptation for MCMC. J. Comp. Graph. Stat. 20:63-79.

Gilks, W.R., Roberts, G.O. and Sahu, S.K. (1998). Adaptive Markov chain Monte Carlo through regeneration. J. Amer. Statist. Assoc. 93:1045-1054

Giordani, P. and Kohn, R. (2010). Adaptive Independent Metropolis-Hastings by Fast Estimation of Mixtures of Normals. J. Comp. Graph. Statist. 19:243-259.

Haario, H., Saksman, E. and Tamminen, J. (2001). An adaptive Metropolis algorithm. Bernoulli 7:223-242

Levine, R. and Casella, G. (2006). Optimizing random scan Gibbs samplers. Journal of Multivariate Analysis 97:2071-2100.

Latuszynski, K. and Roberts, G.O. and Rosenthal, J.S. Adaptive Gibbs samplers and related MCMC methods (2012). Ann. Appl. Prob. (to appear)

Vihola, M. (2012). Robust adaptive Metropolis algorithm with coerced acceptance rate. Statistics and Computing. 22:997-1008.

Adaptive MCMC algorithms (theory)

Andrieu, C. and Moulines, E. (2006). On the ergodicity properties of some adaptive MCMC algorithms. Ann. Appl. Probab. 16:1462-1505

Andrieu, C. and Robert, C.P. (2001). Controlled MCMC for optimal sampling. Technical report, Univ. Paris Dauphine, Ceremade 0125.

Atchadé, Y. and Fort, G. (2010). Limit theorems for some adaptive MCMC algorithms with subgeometric kernels (I). Bernoulli 16:116-154.

Fort, G., Moulines, E. and Priouret, P. (2012). Convergence of interacting MCMC: ergodicity and law of large numbers. Ann. Statist. 39:3262-3289.

Fort, G., Moulines, E., Priouret, P. and Vandekerkhove, P. (2010). Convergence of interacting MCMC: central limit theorem. *submitted* arXiv math.ST 1107-257

Roberts, G.O. and Rosenthal, J.S. (2007). Coupling and ergodicity of adaptive MCMC. J. Appl. Probab. 44:458-475

Saksman, E. and Vihola, M. (2010). On the ergodicity of the adaptive Metropolis algorithm on unbounded domains. Ann. Appl. Probab. 20:2178-2203.

Vihola, M. (2011). On the stability and ergodicity of adaptive scaling Metropolis algorithms. Stochastic Processes and Their Applications, 121:2839-2860.

Interacting samplers : algorithms and theory

Andrieu,, C. and Jasra, A. and Doucet, A. and Del Moral P. (2011) On nonlinear Markov chain Monte Carlo. Bernoulli 17:987-1014.

Chauveau, D. and Vandekerkhove, P. (2001). Improving convergence of the Hastings-Metropolis algorithm with an adaptive proposal. Scand. J. Statist. 29:13-29.

Fort, G., Moulines, E. and Priouret, P. (2012). Convergence of interacting MCMC: ergodicity and law of large numbers. Ann. Statist. 39:3262-3289.

Fort, G., Moulines, E., Priouret, P. and Vandekerkhove, P. (2010). Convergence of interacting MCMC: central limit theorem. *submitted* arXiv math.ST 1107-257

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. Com- puting Science and Statistics: Proc. 23rd Symposium on the Interface, Interface Foundation, Fairfax Station, VA 156-163.

Geyer, C. J. and Thompson, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. J. Am. Statist. Assoc. 90:909-920.

Kou, S., Zhou, Q. and Wong, W. (2006). Equi-energy sampler with applications to statistical inference and statistical mechanisms (with discussion). Ann. Statist. 34:1581-1619.

Marinari, E. and Parisi, G. (1992). Simulated tempering: A new Monte Carlo schemes. Europhysics letters 19:451-458.

Schreck, A. and Fort, G. and Moulines, E. (2012). Adaptive Equi-Energy sampler: Convergence and Illustration. To appear in ACM Transactions on Modeling and Computer Simulation.

Adaptive Importance Sampling

Cappé, O. and Guillin, A. and Marin, J.M. and Robert, C.P. (2004). Population Monte Carlo. J. Comput. Graph. Statist. 13:907-929.

Celeux, G. and Marin, J.M. and Robert, C.P. (2006). Iterated importance sampling in missing data problems. Comput. Statist. Data Anal. 50:3386-3404.

Douc, R. and Guillin, A. and Marin, J.M. and Robert, C.P. (2007). Convergence of Adaptive Mixtures of Importance Sampling Schemes. Ann. Statist. 35:420-448.

Rubinstein, R. Y and Kroese, D. P. (2004). The Cross-Entropy Method: A Unified Approach to Monte Carlo Simulation, Randomized Optimization and Machine Learning. Springer Verlag.

Wraith, D. and Kilbinger, M. and Benabed, K. and Cappé, O. and Cardoso, J.F. and Fort, G. and Prunet, S. and Robert, C.P. (2009). Estimation of cosmological parameters using adaptive importance sampling. Phys. Rev. D, 80

Burn In

Cowles, M.K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. J. Amer. Stat. Assoc. 91:883-904.

Cowles, M.K. and Roberts, G.O. and Rosenthal, J.S. (1999). Possible biases induced by MCMC convergence diagnostics. Journal of Stat. Comput. and Simul. 64:87-104.

Parallelization

Alexopoulos, C. and Andradottir, S. and Argon, N.T. and Goldsman, D. (2006) Replicated batch means variance estimators in the presence of an initial transient. ACM Transactions on Modeling and Computer Simulation, 16:317-328.

Alexopoulos, C. and GOldsman, D. (2004). To batch or not to batch? ACM Transactions on Modeling and Computer Simulation, 14:76-114.

Bratley, P. and Fox, B.L. and Schrage, L.E. (1987) A guide to Simulation. Springer-Verlag, New-York.

Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation uding multiple sequences. Statistical Science, 7:457-472.

Geyer, J.C. (1992) Practical Markov chain Monte Carlo (with discussion). Statistical Science. 7:473-511.

Optimal scaling - to follow

▶ Pioneering work: About the Sym. random walk HM with Gaussian proposal $\mathcal{N}(0,\Gamma)$, in the case

$$\pi(x_1,\cdots,x_d) = \prod_{k=1}^{d} f(x_k) \qquad \Gamma = \frac{s^2}{d} I$$

what is the *optimal* value for s^2?

- Asymptotically, all the components of the chain $X^{(d)}$ are independent and behave as the first one $\{X_k^{(d)}(1), k \geq 0\}$

- Jumps divided by d, so the clock is multiplied by d:

$$Z_t^{(d)} = X_{[td]}^{(d)}(1)$$

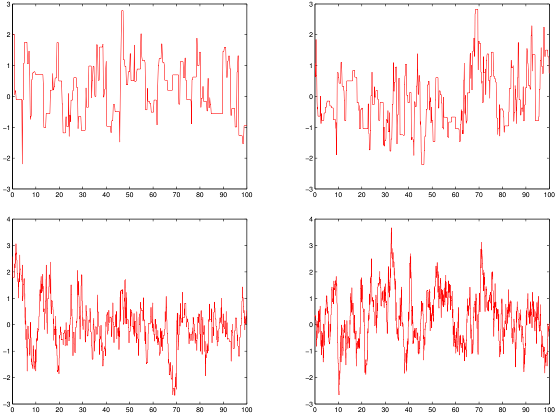
- When $d \to \infty$, $(Z_t^{(d)})_t$ converges to a diffusion process

$$dZ_t = \sqrt{\phi(s)} dB_t + \phi(s) \frac{\nabla \log f(Z_t)}{2} dt.$$

- $\phi(s)$ is the *diffusion coefficient* = speed of the diffusion.

- $s \mapsto \phi(s)$ is optimal at $s = 2.38$.

Optimal scaling - to follow



Skeleton process obtained from a Sym. Random Walk HM chain with target $\mathcal{N}(0,I)$ and proposal $\mathcal{N}(0,\frac{2.38^2}{d}I)$.

In the case $d = 5{,}10$ (top) and $d = 30{,}60$ (bottom).