Limit theorems for adaptive MCMC algorithms

Gersende FORT

LTCI
CNRS - TELECOM ParisTech

In collaboration with Yves ATCHADE (Univ. Michigan, US), Eric MOULINES (TELECOM ParisTech) and Pierre PRIOURET (Univ. Paris 6).

Markov chain Monte Carlo algorithms (MCMC) : algorithms to sample from a **target density** π

- in some applications : known up to a (normalizing) constant.
- complex, so that exact sampling from π is not possible.

Markov chain Monte Carlo algorithms (MCMC) : algorithms to sample from a **target density** π

- in some applications : known up to a (normalizing) constant.
- complex, so that exact sampling from π is not possible.

Define a Markov chain $\{X_n, n \geq 0\}$ with transition kernel : P

$$\mathbb{E}\left[f(X_{n+1})|\mathcal{F}_n\right] = \int f(y)\ P(X_n, dy)$$

so that

- for any bounded function f : $\lim_n \mathbb{E}_x[f(X_n)] = \pi(f)$.
- for any function f increasing like $\cdots$: $n^{-1}\sum_{k=1}^n f(X_k) \longrightarrow_{a.s.} \pi(f)$.
- $\ldots$

I. Adaptive MCMC :

- why ?
- does the process $\{X_n, n \geq 0\}$ approximate π ?

1.1. Symmetric Random Walk Hastings-Metropolis algorithm

An example of transition kernel P is described by the algorithm:

- Choose : a proposal density q
- Iterate: starting from X_n
  - draw (an increment) $Y_{n+1} \sim q(\cdot)$
  - compute the acceptation ratio

$$\alpha(X_n, X_n + Y_{n+1}) := 1 \wedge \frac{\pi(X_n + Y_{n+1})}{\pi(X_n)}$$

  - set

$$X_{n+1} = \begin{cases} Y_{n+1} + X_n & \text{with probability } \alpha(X_n, X_n + Y_{n+1}) \\ X_n & \text{with probability } 1 - \alpha(X_n, X_n + Y_{n+1}) \end{cases}$$

1.1. Symmetric Random Walk Hastings-Metropolis algorithm

An example of transition kernel P is described by the algorithm:

- Choose : a proposal density q
- Iterate: starting from X_n
  - draw (an increment) $Y_{n+1} \sim q(\cdot)$
  - compute the acceptation ratio

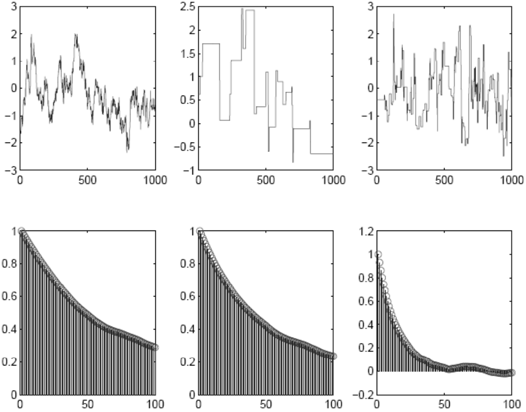
$$\alpha(X_n, X_n + Y_{n+1}) := 1 \wedge \frac{\pi(X_n + Y_{n+1})}{\pi(X_n)}$$

  - set

$$X_{n+1} = \begin{cases} Y_{n+1} + X_n & \text{with probability } \alpha(X_n, X_n + Y_{n+1}) \\ X_n & \text{with probability } 1 - \alpha(X_n, X_n + Y_{n+1}) \end{cases}$$

The efficiency of the algorithm depends upon the proposal q

1.2. On the choice of the variance of the proposal distribution



For ex., when q is Gaussian, how to choose its variance matrix Σ_q ?

- When $\pi \sim \mathcal{N}_d(\mu_\pi, \Sigma_\pi)$, the optimal choice for the variance of q is

$$\Sigma_q = \frac{(2.38)^2}{d} \; \Sigma_\pi .$$

Results obtained by the 'scaling' technique (see also 'fluid limit'). Generalizations exist (other MCMC; relaxing conditions on π)

Roberts-Rosenthal (2001); Bédard (2007); Fort-Moulines-Priouret (2008).

- This suggests an adaptive procedure : learn Σ_π "on the fly" and modify the variance Σ_q continuously during the run of the algorithm.

- When $\pi \sim \mathcal{N}_d(\mu_\pi, \Sigma_\pi)$, the optimal choice for the variance of q is

$$\Sigma_q = \frac{(2.38)^2}{d} \Sigma_\pi .$$

Results obtained by the 'scaling' technique (see also 'fluid limit'). Generalizations exist (other MCMC; relaxing conditions on π)

ROBERTS-ROSENTHAL (2001); BÉDARD (2007); FORT-MOULINES-PRIOURET (2008).

- This suggests an adaptive procedure : learn Σ_π "on the fly" and modify the variance Σ_q continuously during the run of the algorithm.

Example : at each iteration, choose q equal to

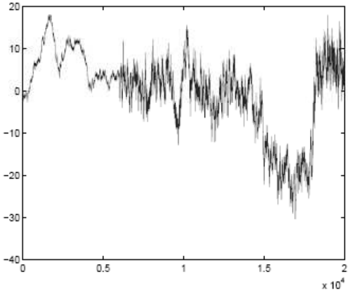
$$0.95\ \mathcal{N}\left(0,(2.38)^2 d^{-1}\ \hat{\Sigma}_n\right) + 0.05\ \mathcal{N}\left(0,(0.1)^2\ d^{-1}\ \mathbb{I}_d\right)$$

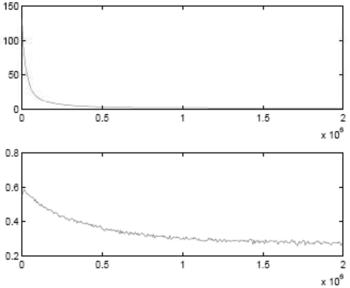
where

$$\hat{\Sigma}_n = \hat{\Sigma}_{n-1} + \frac{1}{n}\left(\{X_n - \mu_n\}\{X_n - \mu_n\}^T - \hat{\Sigma}_{n-1}\right)$$

$$\mu_n = \mu_{n-1} + \frac{1}{n}\left(X_n - \mu_{n-1}\right)$$

HAARIO ET AL. (2001); ROBERTS-ROSENTHAL (2006)


20
10
0
−10
−20
−30
−40
0
0.5
1
1.5
2
x 10^4


150
100
50
0
0
0.5
1
1.5
2
x 10^6
0.8
0.6
0.4
0.2
0
0.5
1
1.5
2
x 10^6

1.3. Be careful with adaptation !

The previous example illustrates the general framework :

- Let $\{P_\theta, \theta \in \Theta\}$ be a family of Markov kernels s.t. $\pi P_\theta = \pi$ for any $\theta \in \Theta$.
- Define a process $\{(\theta_n, X_n), n \geq 0\}$:
  - $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$
  - update θ_{n+1} based on (θ_n, X_n, X_{n+1}) "internal" adaptation

Is it true that the marginal $\{X_n, n \geq 0\}$ approximates π ?

1.3. Be careful with adaptation !

The previous example illustrates the general framework :

- Let $\{P_\theta, \theta \in \Theta\}$ be a family of Markov kernels s.t. $\pi P_\theta = \pi$ for any $\theta \in \Theta$.
- Define a process $\{(\theta_n, X_n), n \geq 0\}$:
  - $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$
  - update θ_{n+1} based on (θ_n, X_n, X_{n+1}) "internal" adaptation

Is it true that the marginal $\{X_n, n \geq 0\}$ approximates π ?

Not always, unfortunately for $\theta \in]0,1[$

$$P_\theta = \begin{bmatrix} (1-\theta) & \theta \\ \theta & (1-\theta) \end{bmatrix} \qquad \pi = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$$

Let $t_1, t_2 \in]0,1[$, and set $\theta_k = t_i$ iff $X_k = i$. Then $\{X_n, n \geq 0\}$ is Markov with invariant probability

$$\tilde{\pi} \propto [t_2 \quad t_1]^T \quad \neq \pi$$

II. Sufficient conditions for convergence of adaptive schemes $\{(\theta_n, X_n), n \geq 0\}$

- convergence of the marginals $\{X_n, n \geq 0\}$
- law of large numbers w.r.t. $\{X_n, n \geq 0\}$

2.1. Convergence of the marginals : Suff Cond

Let

- a family of Markov kernels $\{P_\theta, \theta \in \Theta\}$ s.t. P_θ has an unique invariant probability measure Π_θ
- a filtration $\mathcal{F}_n$ and a process $\{(X_n, \theta_n), n \geq 0\}$ s.t. for any $f \geq 0$,

$$\mathbb{E}\left[f(X_{n+1}) | \mathcal{F}_n\right] = \int f(y) \; P_{\theta_n}(X_n, dy) \qquad \mathbb{P} - a.s.$$

Given a target density $\pi_\star$, which set of conditions will imply

$$\lim_n \sup_{f, |f|_\infty \leq 1} \left|\mathbb{E}\left[f(X_n)\right] - \pi_\star(f)\right| = 0 \quad ?$$

Idea :

$$\mathbb{E}\left[f(X_n)\right] - \pi_\star(f) = \mathbb{E}\left[\mathbb{E}\left[f(X_n)|\mathcal{F}_{n-N}\right]\right] - \pi_\star(f)$$

$$= \mathbb{E}\left[\mathbb{E}\left[f(X_n)|\mathcal{F}_{n-N}\right] - P^N_{\theta_{n-N}} f(X_{n-N})\right] + \mathbb{E}\left[P^N_{\theta_{n-N}} f(X_{n-N}) - \pi_{\theta_{n-N}}(f)\right] + \mathbb{E}\left[\pi_{\theta_{n-N}}(f) - \pi_\star(f)\right]$$

Idea :

$$\mathbb{E}\left[f(X_n)\right] - \pi_\star(f) = \mathbb{E}\left[\mathbb{E}\left[f(X_n)|\mathcal{F}_{n-N}\right]\right] - \pi_\star(f)$$
$$= \mathbb{E}\left[\mathbb{E}\left[f(X_n)|\mathcal{F}_{n-N}\right] - P^N_{\theta_{n-N}} f(X_{n-N})\right] + \mathbb{E}\left[P^N_{\theta_{n-N}} f(X_{n-N}) - \pi_{\theta_{n-N}}(f)\right] + \mathbb{E}\left[\pi_{\theta_{n-N}}(f) - \pi_\star(f)\right]$$

i.e. conditions on

- (Diminishing Adaptation) the difference $\|P_{\theta_n}(x,\cdot) - P_{\theta_{n-1}}(x,\cdot)\|_{\mathrm{TV}}$
- (ergodicity of P_θ / Containment) the convergence of $\|P^N_\theta(x,\cdot) - \pi_\theta\|_{\mathrm{TV}}$ as $N \to +\infty$.
- (convergence of the stationary measures) convergence of $\pi_{\theta_n}(f) - \pi_\star(f)$ as $n \to +\infty$.

Set

$$M_\epsilon(x,\theta) := \inf\{n \geq 1, \|P_\theta^n(x,\cdot) - \pi_\theta\|_{\mathrm{TV}} \leq \epsilon\}.$$

Theorem

Assume

(i) *D.A. cond*	$\sup_x \|P_{\theta_n}(x,\cdot) - P_{\theta_{n-1}}(x,\cdot)\|_{\mathrm{TV}} \longrightarrow_{\mathbb{P}} 0$
(ii) *C. cond*	$\forall \epsilon > 0, \ \lim_M \sup_n \mathbb{P}\left(M_\epsilon(X_n,\theta_n) \geq M\right) = 0$
(iii)	$\pi_\theta = \pi_\star$

Then $\quad \sup_{f, |f|_\infty \leq 1} \left|\mathbb{E}\left[f(X_n)\right] - \pi_\star(f)\right| = 0.$

i.e. conditions on

- (Diminishing Adaptation) the difference $\|P_{\theta_n}(x,\cdot) - P_{\theta_{n-1}}(x,\cdot)\|_{\mathrm{TV}}$
- (ergodicity of P_θ / Containment) the convergence of $\|P_\theta^N(x,\cdot) - \pi_\theta\|_{\mathrm{TV}}$ as $N \to +\infty$.
- (convergence of the stationary measures) convergence of $\pi_{\theta_n}(f) - \pi_\star(f)$ as $n \to +\infty$.

Set

$$M_\epsilon(x,\theta) := \inf\{n \geq 1, \|P_\theta^n(x,\cdot) - \pi_\theta\|_{\mathrm{TV}} \leq \epsilon\}.$$

Theorem

Assume

(i) *D.A. cond* $\sup_x \|P_{\theta_n}(x,\cdot) - P_{\theta_{n-1}}(x,\cdot)\|_{\mathrm{TV}} \longrightarrow_{\mathbb{P}} 0$

(ii) *C. cond* $\forall \epsilon > 0, \; \lim_M \sup_n \mathbb{P}\left(M_\epsilon(X_n,\theta_n) \geq M\right) = 0$

(iii) $\forall \epsilon > 0, \; \sup_{f\in\mathsf{F}} \mathbb{P}\left(|\pi_{\theta_n}(f) - \pi_\star(f)| > \epsilon\right) \to 0$

Then $\sup_{f\in\mathsf{F}} \left|\mathbb{E}\left[f(X_n)\right] - \pi_\star(f)\right| = 0.$

i.e. conditions on

- (Diminishing Adaptation) the difference $\|P_{\theta_n}(x,\cdot) - P_{\theta_{n-1}}(x,\cdot)\|_{\mathrm{TV}}$
- (ergodicity of P_θ / Containment) the convergence of $\|P_\theta^N(x,\cdot) - \pi_\theta\|_{\mathrm{TV}}$ as $N \to +\infty$.
- (convergence of the stationary measures) convergence of $\pi_{\theta_n}(f) - \pi_\star(f)$ as $n \to +\infty$.

2.2. Convergence of the marginals : in 'practice'

It is sufficient to establish

- (D.A. cond) problem specific
- (C. cond) a uniform-in-θ drift condition (geometric or sub-geometric drift) and a uniform-in-θ minorization of the transition kernel

  (Roberts-Rosenthal (2007); Bai (2009); Atchadé-Fort (2009))
- (Cvg π_{θ_n}) $\exists \theta_\star$ and a set A s.t. $\mathbb{P}(A) = 1$ and

$$\forall \omega \in A, \qquad \forall x, \forall B \qquad\qquad P_{\theta_n(\omega)}(x,B) = P_{\theta_\star}(x,B).$$

3.1. Strong law of large numbers : Suff cond

Let

- a family of Markov kernels $\{P_\theta, \theta \in \Theta\}$ s.t. P_θ has an unique invariant probability measure π_θ
- a filtration $\mathcal{F}_n$ and a process $\{(X_n, \theta_n), n \geq 0\}$ s.t. for any $f \geq 0$,

$$\mathbb{E}\left[f(X_{n+1}) | \mathcal{F}_n\right] = \int f(y) \; P_{\theta_n}(X_n, dy) \qquad \mathbb{P}\text{-a.s.}$$

Given a target density $\pi_\star$, which set of conditions will imply

$$n^{-1} \sum_{k=1}^{n} f(X_k) \to \pi_\star(f) \qquad \mathbb{P} - a.s.$$

for a large class of functions f ?

Idea :

$$n^{-1} \sum_{k=1}^{n} f(X_k) - \pi_\star(f)$$
$$= n^{-1} \sum_{k=1}^{n} \{f(X_k) - \pi_{\theta_{k-1}}(f)\} + n^{-1} \sum_{k=0}^{n-1} \{\pi_{\theta_k}(f) - \pi_\star(f)\}$$
$$= M_n(f) + R_n(f) + n^{-1} \sum_{k=0}^{n-1} \{\pi_{\theta_k}(f) - \pi_\star(f)\}$$

where M_n : martingale.

Idea :

$$n^{-1}\sum_{k=1}^{n} f(X_k) - \pi_\star(f)$$
$$= n^{-1}\sum_{k=1}^{n}\{f(X_k) - \pi_{\theta_{k-1}}(f)\} + n^{-1}\sum_{k=0}^{n-1}\{\pi_{\theta_k}(f) - \pi_\star(f)\}$$
$$= M_n(f) + R_n(f) + n^{-1}\sum_{k=0}^{n-1}\{\pi_{\theta_k}(f) - \pi_\star(f)\}$$

where M_n : martingale.

i.e. conditions for

- a.s. conv for martingales : from conditions on L^p-moments for the increments $(p > 1)$.
- a.s. conv of the residual terms : from a strenghtened diminishing adaptation condition ($\longleftrightarrow$ conditions on the regularity in θ of the Poisson equation)
- a.s. conv of the stationary measures : from the "a.s." conv of $P_{\theta_n}(x,B)$ to $P_{\theta_\star}(x,B)$

3.2. Strong law of large numbers "in practice"

It is sufficient to establish

- (strenghtened D.A. cond) problem specific
- (C. cond) a uniform-in-θ drift condition (geometric or sub-geometric drift) and a uniform-in-θ minorization of the transition kernel

  (Roberts-Rosenthal (2007); Bai (2009); Atchadé-Fort (2009))
- (Cvg π_{θ_n}) $\exists \theta_\star$ and a set A s.t. $\mathbb{P}(A) = 1$ and

$$\forall \omega \in A, \qquad \forall x, \forall B \qquad\qquad P_{\theta_n(\omega)}(x,B) = P_{\theta_\star}(x,B).$$

3.2. Strong law of large numbers “in practice”

It is sufficient to establish

- (strenghtened D.A. cond) problem specific
- (C. cond) a uniform-in-θ drift condition (geometric or sub-geometric drift) and a uniform-in-θ minorization of the transition kernel

(Roberts-Rosenthal (2007); Bai (2009); Atchadé-Fort (2009))

- (Cvg π_{θ_n}) $\exists \theta_\star$ and a set A s.t. $\mathbb{P}(A) = 1$ and

$$\forall \omega \in A, \qquad \forall x, \forall B \qquad\qquad P_{\theta_n(\omega)}(x,B) = P_{\theta_\star}(x,B).$$

When the drift condition is of the form :

- (Geom) $P_\theta V \leq \lambda V + b \mathbb{1}_{\mathcal{C}}$: strong law of large numbers for functions increasing like V^α for any $\alpha \in [0,1[$.
- (Sub-Geom) $P_\theta V \leq V - c\, V^{1-\alpha} + b\mathbb{1}_{\mathcal{C}}$: strong law of large numbers for functions increasing like V^β for any $\beta \in [0,1-\alpha[$.

Conclusion

We provide answers to the problem : given

- a family of Markov kernels $\{P_\theta, \theta \in \Theta\}$ s.t. P_θ has an unique invariant probability distribution π_θ
- a filtration $\mathcal{F}_n$ and a process $\{(X_n, \theta_n), n \geq 0\}$ s.t. for any $f \geq 0$,

$$\mathbb{E}\left[f(X_{n+1}) | \mathcal{F}_n\right] = \int f(y) \; P_{\theta_n}(X_n, dy) \qquad \mathbb{P}\text{-a.s.}$$

- which set of conditions will imply
  - convergence of the distribution of $\{X_n, n \geq 0\}$ to some prob. $\pi_\star$
  - convergence of the empirical distribution $n^{-1} \sum_{k=1}^n \delta_{X_k}$
- Appli: convergence of “internal” and “external” adaptive MCMC.

- Details in
  - **Y. Atchadé, G. Fort** Limit theorems for some adaptive MCMC algorithms with subgeometric kernels, *Accepted in Bernoulli, 2009*
  - **Y. Atchadé, G. Fort, E. Moulines, P. Priouret** Adaptive MCMC : theory and practice, *submitted*