Convergence of Adaptive and Interacting MCMC algorithms

Gersende FORT

LTCI / CNRS - TELECOM ParisTech, France

Joint work with E. Moulines (LTCI, France) and P. Priouret (LPMA, France)

Examples of adaptive MCMC

Convergence of the marginals for adaptive MCMC samplers

Law of large numbers for adaptive MCMC samplers

Convergence of the stationary distributions π_{θ_n}

Applications

I. Two examples of adaptive MCMC samplers

1. an Adaptive MCMC algorithm
2. an Interacting MCMC algorithm

Example 1: The Adaptive Metropolis

[Haario et al. (1999)]

Consider the Metropolis-Hastings algorithm

- with target density π on X — $\mathsf{X} \subseteq \mathbb{R}^d$, density w.r.t. the Lebesgue measure
- with Gaussian proposal $q_{\theta}(x, y) = \mathcal{N}_d(x, \theta)[y]$

$\hookrightarrow$ How to choose the design parameter θ ?

Example 1: The Adaptive Metropolis

[HAARIO ET AL. (1999)]

Consider the Metropolis-Hastings algorithm

- with target density π on X   X $\subseteq \mathbb{R}^d$, density w.r.t. the Lebesgue measure
- with Gaussian proposal $q_{\theta}(x, y) = \mathcal{N}_d(x, \theta)[y]$

$\hookrightarrow$ How to choose the design parameter θ ?

Ans: covariance matrix of π up to a scalar, [ROBERTS ET AL. (1997)] iteratively estimated by the empirical covariance matrix or a robust estimator

$$\theta_{n+1} = \frac{n}{n+1}\theta_n + \frac{1}{n+1}\left\{(X_{n+1} - \mu_{n+1})(X_{n+1} - \mu_{n+1})^T + \kappa \ \mathrm{Id}_d\right\}$$

$$\mu_{n+1} = \mu_n + \frac{1}{n+1}(X_{n+1} - \mu_n)$$

This yields the adaptive Metropolis algorithm: iteratively

- draw $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$ transition kernel of a HM algo with Gaussian proposal with covariance matrix $\propto \theta_n$
- update the parameter θ_{n+1}, based on θ_n and $X_{1:n+1}$

This yields the adaptive Metropolis algorithm: iteratively

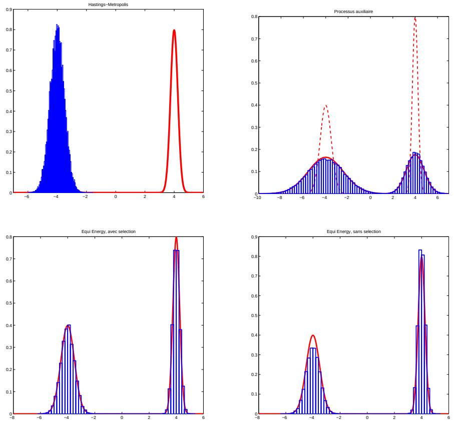
- draw $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$ transition kernel of a HM algo with Gaussian proposal with covariance matrix $\propto \theta_n$
- update the parameter θ_{n+1}, based on θ_n and $X_{1:n+1}$

In this example

- $\pi P_\theta = \pi$ i.e. same invariant distribution
- $\theta_n \in \Theta$ where Θ is a finite dimensional parameter space.

Example 2: The Equi-Energy sampler (simplified) [Kou et al. (2006)]

$\hookrightarrow$ For the simulation of multi-modal density π.



Let

- a transition kernel P such that $\pi P = \pi$.
- a probability of swap: $\epsilon \in (0,1)$
- an auxiliary process $\{Y_n, n \geq 0\}$ that “targets” the tempered density $\pi^{1-\beta}$ $\quad (\beta \in (0,1))$

Let

- a transition kernel P such that $\pi P = \pi$.
- a probability of swap: $\epsilon \in (0,1)$
- an auxiliary process $\{Y_n, n \geq 0\}$ that "targets" the tempered density $\pi^{1-\beta}$ $(\beta \in (0,1))$

Define iteratively the process of interest $\{X_n, n \geq 0\}$

- with probability $(1-\epsilon)$: draw $X_{n+1} \sim P(X_n, \cdot)$
- with probability ϵ: draw at random Y through the past values $Y_{0:n}$ and accept or not Y as the new value, with an acceptation-rejection algorithm.

Let

- a transition kernel P such that $\pi P = \pi$.
- a probability of swap: $\epsilon \in (0,1)$
- an auxiliary process $\{Y_n, n \geq 0\}$ that "targets" the tempered density $\pi^{1-\beta}$ $(\beta \in (0,1))$

Define iteratively the process of interest $\{X_n, n \geq 0\}$

- with probability $(1-\epsilon)$: draw $X_{n+1} \sim P(X_n, \cdot)$
- with probability ϵ: draw at random Y through the past values $Y_{0:n}$ and accept or not Y as the new value, with an acceptation-rejection algorithm. (simplified EE)

This yields the (simplified) Equi-Energy sampler: $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$

where $$\theta_{n+1} = \frac{1}{n+1} \sum_{k=0}^{n} \delta_{Y_k}$$

$$P_\theta(x, A) = (1-\epsilon) P(x, A) + \epsilon \left\{ \int_A \alpha(x,y) \theta(\mathrm{d}y) + \mathbb{1}_A(x) \int (1-\alpha(x,y)) \theta(\mathrm{d}y) \right\}$$

and $\alpha(x, y)$ defined such that $\pi P_{\theta_\star} = \pi$ where $\theta_\star = \lim_n \theta_n \propto \pi^{1-\beta}$

This yields the (simplified) Equi-Energy sampler: $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$

where $$\theta_{n+1} = \frac{1}{n+1} \sum_{k=0}^{n} \delta_{Y_k}$$

$$P_\theta(x, A) = (1-\epsilon)P(x, A) + \epsilon \left\{ \int_A \alpha(x,y)\theta(\mathrm{d}y) + \mathbb{1}_A(x) \int (1-\alpha(x,y))\theta(\mathrm{d}y) \right\}$$

and $\alpha(x, y)$ defined such that $\pi P_{\theta_\star} = \pi$ where $\theta_\star = \lim_n \theta_n \propto \pi^{1-\beta}$

In this example

- $\pi_\theta P_\theta = \pi_\theta$ i.e. invariant distributions depending upon θ
- $\theta_n \in \Theta$ where Θ is an infinite dimensional parameter space.

Conclusion

Let a family of transition kernels on X, $\{P_\theta, \theta \in \Theta\}$.

Consider a X $\times$ Θ-valued process $\{(X_n, \theta_n), n \geq 0\}$ such that

- it is adapted to a filtration $\mathcal{F}_n$.
- $\mathbb{P}(X_{n+1} \in A | \mathcal{F}_n) = P_{\theta_n}(X_n, A)$

$\hookrightarrow$ What kind of conditions on the adaption mecanism $\{\theta_n, n \geq 0\}$ and on the transition kernels $\{P_\theta, \theta \in \Theta\}$ imply that there exists a distribution π such that

- convergence of the marginals: $\mathbb{E}\left[f(X_n)\right] \to \pi(f)$ $\quad$ f bounded
- law of large numbers: $n^{-1} \sum_{k=1}^{n} f(X_k) \xrightarrow{\text{a.s.}} \pi(f)$ $\quad$ f unbounded

II. Convergence of the marginals for adaptive MCMC samplers

For a bounded function f,

$$\mathbb{E}\left[f(X_n)\right] - \pi(f) \to 0$$

Even in the case the kernels P_θ have <u>*the same*</u> *invariant distribution, it is NOT true that ergodicity holds since the kernels are chosen at random. Conditions on the adaptation mecanism are required*

Sketch of the proof

We write

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

Sketch of the proof

We write

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

$\hookrightarrow$ [A] condition on the ergodicity of the transition kernels

"Usually", the transition kernels $\{P_\theta, \theta \in \Theta\}$ are geometrically ergodic :

$$\sup_{f, |f| \leq 1} |P^n_\theta f(x) - \pi_\theta(f)| \leq C_\theta \ \rho^n_\theta \ V(x) \qquad \rho_\theta \in (0,1)$$

Sketch of the proof

We write

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

$\hookrightarrow$ [B] condition on the adaptation mecanism since

$$\left|\mathbb{E}\left[f(X_n) - P^N_{\theta_{n-N}} f(X_{n-N})\right]\right| \leq \sum_{j=1}^{N-1} (N-j)\mathbb{E}\left[\sup_x \left\| P_{\theta_{n-N+j}}(x,\cdot) - P_{\theta_{n-N+j-1}}(x,\cdot)\right\|_{\mathrm{TV}}\right]$$

Sketch of the proof

We write

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

$\hookrightarrow$ [C] when $\pi_\theta \neq \pi$, condition on the convergence of $\{\pi_{\theta_n}, n \geq 0\}$ to π

Sketch of the proof

We write

$$\mathbb{E}\left[f(X_n)\right] - \pi(f) = \mathbb{E}\left[f(X_n) - P^{r(n)}_{\theta_{n-r(n)}} f(X_{n-r(n)})\right] + \mathbb{E}\left[P^{r(n)}_{\theta_{n-r(n)}} f(X_{n-r(n)}) - \pi_{\theta_{n-r(n)}}(f)\right] + \mathbb{E}\left[\pi_{\theta_{n-r(n)}}(f)\right] - \pi(f)$$

The conditions can be weakened by replacing N by $r(n)$. This allows to consider situations when the *transition kernels are not simultaneously ergodic*

$$\sup_{f,|f|\leq 1} |P^n_{\theta} f(x) - \pi_{\theta}(f)| \leq C_{\theta} \; \rho^n_{\theta} \; V(x) \qquad \rho_{\theta} \in (0,1)$$

and even cases where $C_{\theta_n} \vee (1-\rho_{\theta_n})^{-1}$ is not bounded (a.s.).

Result [Fort et al. 2010]

A. (Ergodicity of the transition kernels)

- $\exists \pi_\theta$ s.t. $\pi_\theta P_\theta = \pi_\theta$
- for any $\epsilon > 0$, there exists a non-decreasing positive sequence $\{r_\epsilon(n), n \geq 0\}$ such that $\limsup_{n\to\infty} r_\epsilon(n)/n = 0$ and

$$\limsup_{n\to\infty} \mathbb{E}\left[\left\| P^{r_\epsilon(n)}_{\theta_{n-r_\epsilon(n)}}(X_{n-r_\epsilon(n)}, \cdot) - \pi_{\theta_{n-r_\epsilon(n)}} \right\|_{\mathrm{TV}} \right] \leq \epsilon \; .$$

B. (Diminishing adaptation) For any $\epsilon > 0$,

$$\lim_{n\to\infty} \sum_{j=0}^{r_\epsilon(n)-1} \mathbb{E}\left[\sup_x \left\| P_{\theta_{n-r_\epsilon(n)+j}}(x, \cdot) - P_{\theta_{n-r_\epsilon(n)}}(x, \cdot) \right\|_{\mathrm{TV}} \right] = 0$$

C. (Convergence of the invariant distributions) $\exists\ \pi$ and a bounded non-negative function f s.t. $\lim_n \pi_{\theta_n}(f) = \pi(f)$ a.s.

Then $\lim_n \mathbb{E}\left[f(X_n) \right] = \pi(f)$.

Comparison with the literature pioneering work by [Roberts & Rosenthal, 2007]

1. Our conditions both weaken the *containment condition* and the *diminishing adaptation condition* of [Roberts & Rosenthal, 2007]. We are able to consider cases when the transition kernels are ergodic but not necessarily uniformly-in-θ.

$$\sup_{f, |f| \leq 1} |P_\theta^n f(x) - \pi_\theta(f)| \leq C_\theta \; \rho_\theta^n \; V(x)$$

Nevertheless, it is required to have an explicit control of ergodicity s.t. $C_{\theta_n} \vee (1-\rho_{\theta_n})^{-1}$ does not "explode too rapidly".

Comparison with the literature pioneering work by [Roberts & Rosenthal, 2007]

1. Our conditions both weaken the *containment condition* and the *diminishing adaptation condition* of [Roberts & Rosenthal, 2007]. We are able to consider cases when the transition kernels are ergodic but not necessarily uniformly-in-θ.

$$\sup_{f,|f|\leq 1} |P_\theta^n f(x) - \pi_\theta(f)| \leq C_\theta \; \rho_\theta^n \; V(x)$$

   Nevertheless, it is required to have an explicit control of ergodicity s.t. $C_{\theta_n} \vee (1-\rho_{\theta_n})^{-1}$ does not "explode too rapidly".

2. π_θ can depend upon θ provided we are able to prove that $\pi_{\theta_n}(f)$ converges to $\pi(f)$.

III. Law of large numbers for adaptive MCMC samplers

For an (unbounded) function f s.t. $\cdots$

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \xrightarrow{\text{a.s.}} \pi(f).$$

Sketch of the proof

We write

$$n^{-1}\sum_{k=1}^{n} f(X_k) - \pi(f) = n^{-1}\sum_{k=1}^{n}\left\{f(X_k) - \pi_{\theta_{k-1}}(f)\right\} + \frac{1}{n}\sum_{k=1}^{n}\pi_{\theta_{k-1}}(f) - \pi(f)$$

For the second term, $\hookrightarrow$ [A] condition on $\pi_{\theta_n}(f) \xrightarrow{\text{a.s.}} \pi(f)$

Sketch of the proof

$$n^{-1}\sum_{k=1}^{n} f(X_k) - \pi(f) = n^{-1}\sum_{k=1}^{n}\left\{f(X_k) - \pi_{\theta_{k-1}}(f)\right\} + \frac{1}{n}\sum_{k=1}^{n}\pi_{\theta_{k-1}}(f) - \pi(f)$$

For the first term, Tool : Poisson equation so that

$$n^{-1}\sum_{k=1}^{n}\left\{f(X_k) - \pi_{\theta_{k-1}}(f)\right\} = n^{-1}\underbrace{\sum_{k=1}^{n}\Delta M_k}_{\text{sum of martingale increments}} + \underbrace{R_n^{(1)}}_{\text{Rest due to the adaptation}} + \underbrace{R_n^{(2)}}_{\text{Rest}}$$

Sketch of the proof

$$n^{-1}\sum_{k=1}^{n} f(X_k) - \pi(f) = n^{-1}\sum_{k=1}^{n}\left\{f(X_k) - \pi_{\theta_{k-1}}(f)\right\} + \frac{1}{n}\sum_{k=1}^{n}\pi_{\theta_{k-1}}(f) - \pi(f)$$

For the first term, Tool : Poisson equation so that

$$n^{-1}\sum_{k=1}^{n}\left\{f(X_k) - \pi_{\theta_{k-1}}(f)\right\} = n^{-1} \underbrace{\sum_{k=1}^{n}\Delta M_k}_{\text{sum of martingale increments}} + \underbrace{R_n^{(1)}}_{\text{Rest due to the adaptation}} + \underbrace{R_n^{(2)}}_{\text{Rest}}$$

- Martingale increments : $\hookrightarrow$ [B] moment conditions of the form

$$\exists \alpha > 1, \qquad \sum_k \frac{1}{k^\alpha}\, \mathbb{E}\left[|\Delta M_k|^\alpha | \mathcal{F}_{k-1}\right] < +\infty \quad \text{a.s.}$$

Sketch of the proof

$$n^{-1}\sum_{k=1}^{n} f(X_k)-\pi(f) = n^{-1}\sum_{k=1}^{n}\left\{f(X_k)-\pi_{\theta_{k-1}}(f)\right\}+\frac{1}{n}\sum_{k=1}^{n}\pi_{\theta_{k-1}}(f)-\pi(f)$$

For the first term, Tool : Poisson equation so that

$$n^{-1}\sum_{k=1}^{n}\left\{f(X_k)-\pi_{\theta_{k-1}}(f)\right\} = n^{-1}\underbrace{\sum_{k=1}^{n}\Delta M_k}_{\text{sum of martingale increments}} + \underbrace{R_n^{(1)}}_{\text{Rest due to the adaptation}} + \underbrace{R_n^{(2)}}_{\text{Rest}}$$

- $R_n^{(1)}$: $\hookrightarrow$ [C] condition on the adaptation: “diminishing adaptation”
- $R_n^{(2)}$: $\hookrightarrow$ very weak conditions ! (more or less, a consequence of the other conditions).

Result

[Fort et al. 2010]

A. (Ergodicity of the transition kernels) There exist $C_\theta, \rho_\theta \in (0,1)$ s.t.

$$\|P_\theta^n(x,\cdot) - \pi_\theta\|_V \leq C_\theta \; \rho_\theta^n \; V(x)$$

B. (Martingale term) $\exists \alpha > 1$

$$\sum_k \frac{1}{k^\alpha} \; \left(C_{\theta_k} \vee (1-\rho_{\theta_k})^{-1}\right)^{2\alpha} \; P_{\theta_k} V^\alpha(X_k) < +\infty \;\; \text{a.s.}$$

C. (Strenghtened diminishing adaptation)

$$\sum_k \frac{1}{k} \left(C_{\theta_k} \vee (1-\rho_{\theta_k})^{-1}\right)^6 \; V(X_k) \; \sup_x \; \sup_{f, |f| \leq V} \frac{|P_{\theta_k} f(x) - P_{\theta_{k-1}} f(x)|}{V(x)} < \infty \;\; \text{a.s.}$$

D. (Convergence of the inv. distributions) for f s.t. $|f| \leq V^a, a \in (0,1)$

$$\pi_{\theta_n}(f) \xrightarrow{\text{a.s.}} \pi(f)$$

Then, $\quad n^{-1} \sum_{k=1}^n f(X_k) \xrightarrow{\text{a.s.}} \pi(f)$

Comparison with the literature

[Atchadé & Rosenthal (2005), Andrieu & Moulines (2006), Roberts & Rosenthal (2007), Saksman & Vihola (2008), Vihola (2009), Atchadé & Fort (2010), Atchad et al. (2010) $\cdots$]

- We are able to prove a strong law of large numbers, for unbounded functions
- without assuming a uniform-in-θ ergodic behavior on the transition kernels (neither the state space X nor the parameter space Θ have to be compact/countable/finite)
- under the condition that the adaptation is diminishing which does not require that the sequence $\{\theta_n, n \geq 0\}$ converges (for example, adaptation based on a stochastic approximation dynamic: "$\theta_n = \theta_{n-1} + \gamma_n H_n(\theta_n, W_{n+1})$" is OK)
- without assuming the stability of the sequence $\{\theta_n, n \geq 0\}$ for example in the finite dimensional case, control of the form "$\limsup_n n^{-\tau} |\theta_n| < +\infty$ a.s. for $\tau > 0$" is OK (at least when $\pi_\theta = \pi \cdots$ - see next section)

IV. Convergence of the stationary distributions

Under the (main) assumption There exists $\theta_\star$ s.t. for any $x \in \mathsf{X}, A \in \mathcal{X}$

$$\exists \Omega_{x,A}, \qquad \mathbb{P}(\Omega_{x,A}) = 1 \qquad \forall \omega \in \Omega_{x,A} \qquad \lim_n P_{\theta_n(\omega)}(x,A) = P_{\theta_\star}(x,A)$$

we prove that for any bounded and continuous function f,

$$\exists \Omega_\star, \qquad \mathbb{P}(\Omega_\star) = 1 \qquad \forall \omega \in \Omega_\star \qquad \lim_n \pi_{\theta_n(\omega)}(f) = \pi_{\theta_\star}(f) \; .$$

well, we have even a stronger result, $\Omega_\star$ does not depend upon f

We write

$$\pi_{\theta_n}(f) - \pi_{\theta_\star}(f) = \left(\pi_{\theta_n}(f) - P^k_{\theta_n} f(x)\right) + \left(P^k_{\theta_n} f(x) - P^k_{\theta_\star} f(x)\right) + \left(P^k_{\theta_\star} f(x) - \pi_{\theta_\star}(f)\right)$$

and control the blue terms by a condition on the ergodicity of the transition kernels.

We write

$$\pi_{\theta_n}(f) - \pi_{\theta_\star}(f) = \left(\pi_{\theta_n}(f) - P^k_{\theta_n} f(x)\right) + \left(P^k_{\theta_n} f(x) - P^k_{\theta_\star} f(x)\right) + \left(P^k_{\theta_\star} f(x) - \pi_{\theta_\star}(f)\right)$$

and control the blue terms by a condition on the ergodicity of the transition kernels.

For the control of the red term, we write

$$P^k_{\theta_n} f(x) - P^k_{\theta_\star} f(x) = \int \left(P_{\theta_n}(x, \mathrm{d}y) - P_{\theta_\star}(x, \mathrm{d}y)\right) P^{k-1}_{\theta_\star} f(y) + \int P_{\theta_n}(x, \mathrm{d}y) \left(P^{k-1}_{\theta_n} f(y) - P^{k-1}_{\theta_\star} f(y)\right)$$

Starting from :

$$\forall x \in \mathsf{X}, A \in \mathcal{X}, \quad \exists \Omega_{x,A}, \quad \mathbb{P}(\Omega_{x,A}) = 1 \quad \forall \omega \in \Omega_{x,A} \quad \lim_n P_{\theta_n(\omega)}(x, A) = P_{\theta_\star}(x, A)$$

Starting from :

$$\forall x \in \mathsf{X}, A \in \mathcal{X}, \quad \exists \Omega_{x,A}, \quad \mathbb{P}(\Omega_{x,A}) = 1 \quad \forall \omega \in \Omega_{x,A} \quad \lim_n P_{\theta_n(\omega)}(x, A) = P_{\theta_\star}(x, A)$$

the steps are:

$$\forall x \in \mathsf{X}, \quad \exists \Omega_x, \qquad \mathbb{P}(\Omega_x) = 1 \qquad \forall \omega \in \Omega_x \qquad \lim_n P_{\theta_n(\omega)}(x, \cdot) \xrightarrow{w} P_{\theta_\star}(x, \cdot)$$

$\hookrightarrow$ Tool: separable metric space X (ex. Polish)

Starting from :

$$\forall x \in \mathsf{X}, A \in \mathcal{X}, \quad \exists \Omega_{x,A}, \quad \mathbb{P}(\Omega_{x,A}) = 1 \quad \forall \omega \in \Omega_{x,A} \quad \lim_n P_{\theta_n(\omega)}(x, A) = P_{\theta_\star}(x, A)$$

the steps are:

$$\forall x \in \mathsf{X}, \quad \exists \Omega_x, \qquad \mathbb{P}(\Omega_x) = 1 \qquad \forall \omega \in \Omega_x \qquad \lim_n P_{\theta_n(\omega)}(x, \cdot) \xrightarrow{w} P_{\theta_\star}(x, \cdot)$$

$\hookrightarrow$ Tool: separable metric space X (ex. Polish)

$$\exists \Omega', \qquad \mathbb{P}(\Omega') = 1 \qquad \forall \omega \in \Omega', x \in \mathsf{X} \qquad \lim_n P_{\theta_n(\omega)}(x, \cdot) \xrightarrow{w} P_{\theta_\star}(x, \cdot) \ ,$$

$\hookrightarrow$ Tool: Polish space X + equicontinuity of $\{P_\theta f - P_{\theta_\star} f, \theta \in \Theta\}$

Starting from :

$$\forall x \in \mathsf{X}, A \in \mathcal{X}, \quad \exists \Omega_{x,A}, \quad \mathbb{P}(\Omega_{x,A}) = 1 \quad \forall \omega \in \Omega_{x,A} \quad \lim_n P_{\theta_n(\omega)}(x, A) = P_{\theta_\star}(x, A)$$

the steps are:

$$\forall x \in \mathsf{X}, \quad \exists \Omega_x, \qquad \mathbb{P}(\Omega_x) = 1 \qquad \forall \omega \in \Omega_x \qquad \lim_n P_{\theta_n(\omega)}(x, \cdot) \xrightarrow{w} P_{\theta_\star}(x, \cdot)$$

$\hookrightarrow$ Tool: separable metric space X (ex. Polish)

$$\exists \Omega', \qquad \mathbb{P}(\Omega') = 1 \qquad \forall \omega \in \Omega', x \in \mathsf{X} \qquad \lim_n P_{\theta_n(\omega)}(x, \cdot) \xrightarrow{w} P_{\theta_\star}(x, \cdot) \ ,$$

$\hookrightarrow$ Tool: Polish space X + equicontinuity of $\{P_\theta f - P_{\theta_\star} f, \theta \in \Theta\}$

$$\exists \Omega_\star, \qquad \mathbb{P}(\Omega_\star) = 1 \qquad \forall \omega \in \Omega_\star \qquad \lim_n P^k_{\theta_n(\omega)}(x, \cdot) \xrightarrow{w} P^k_{\theta_\star}(x, \cdot) \ ,$$

$\hookrightarrow$ Tool: Feller properties of the kernels $\{P_\theta, \theta \in \Theta\}$

Result

[FORT ET AL. 2010]

A. (Ergodicity of the transition kernels)

B. X is Polish

C. $P_{\theta_\star}$ is Feller and for any bounded continuous function f,
$\{P_\theta f, \theta \in \Theta\}$ is equicontinuous.

D. (Convergence of the transition kernels) for any $x \in \mathsf{X}$,
$P_{\theta_n}(x, \cdot) \xrightarrow{w} P_{\theta_\star}(x, \cdot)$ a.s..

Then for any bounded continuous function f, $\pi_{\theta_n}(f) \xrightarrow{\text{a.s.}} \pi_{\theta_\star}(f)$.

Result [Fort et al. 2010]

A. (Ergodicity of the transition kernels)

B. X is Polish

C. $P_{\theta_\star}$ is Feller and for any bounded continuous function f, $\{P_\theta f, \theta \in \Theta\}$ is equicontinuous.

D. (Convergence of the transition kernels) for any $x \in \mathsf{X}$, $P_{\theta_n}(x, \cdot) \xrightarrow{w} P_{\theta_\star}(x, \cdot)$ a.s..

Then for any bounded continuous function f, $\pi_{\theta_n}(f) \xrightarrow{\text{a.s.}} \pi_{\theta_\star}(f)$.

Rmk: Extensions to unbounded continuous functions by (standard) moment conditions.

V. Application to the convergence of adaptive and interacting MCMC algorithms

Ergodicity criteria: checked in practice by

- drift inequality $P_\theta V \leq \lambda_\theta V + b_\theta$
- minorization condition $P_\theta(x, \cdot) \geq \delta_\theta \; \nu_\theta(\cdot) \mathbb{1}_{\mathcal{C}_\theta}(x)$
- conditions on the decay of the rate ξ s.t.

  $\limsup_n \xi(n) \; \left(b_{\theta_n} \vee \delta_{\theta_n}^{-1} \vee (1-\lambda_{\theta_n})^{-1}\right) < +\infty$

V. Application to the convergence of adaptive and interacting MCMC algorithms

Ergodicity criteria: checked in practice by

- drift inequality $P_\theta V \leq \lambda_\theta V + b_\theta$
- minorization condition $P_\theta(x, \cdot) \geq \delta_\theta \; \nu_\theta(\cdot) \mathbb{1}_{\mathcal{C}_\theta}(x)$
- conditions on the decay of the rate ξ s.t.

  $\limsup_n \xi(n) \; \left(b_{\theta_n} \vee \delta_{\theta_n}^{-1} \vee (1-\lambda_{\theta_n})^{-1}\right) < +\infty$

Diminishing adaptation: checked in practice by

$$\mathrm{distance}(P_\theta, P_{\theta'}) \leq C \; \mathrm{distance}(\theta, \theta') \qquad \text{for some "distance"}$$

V. Application to the convergence of adaptive and interacting MCMC algorithms

Ergodicity criteria: checked in practice by

- drift inequality $P_\theta V \leq \lambda_\theta V + b_\theta$
- minorization condition $P_\theta(x, \cdot) \geq \delta_\theta \ \nu_\theta(\cdot) \mathbb{1}_{\mathcal{C}_\theta}(x)$
- conditions on the decay of the rate ξ s.t.
  $\limsup_n \xi(n) \ \left(b_{\theta_n} \vee \delta_{\theta_n}^{-1} \vee (1-\lambda_{\theta_n})^{-1} \right) < +\infty$

Diminishing adaptation: checked in practice by

$$\mathrm{distance}(P_\theta, P_{\theta'}) \leq C \ \mathrm{distance}(\theta, \theta') \qquad \text{for some "distance"}$$

Convergence of $\{\pi_{\theta_n}(f), n \geq 0\}$ when $\pi_\theta \neq \pi$: based on the convergence of $\{\theta_n, n \geq 0\}$

Adaptive MCMC

We prove

- when the target density π is *lighter than exponential*
- with $\mathcal{N}_d$ (adapted) proposal distribution s.t. the eigenvalues of the cov matrix are larger than κ.

1. Ergodicity: $\lim_n \sup_{f,|f|_\infty \leq 1} \mathbb{E}\left[f(X_n)\right] = \pi(f)$. contemporaneous work by (Bai et al., 2010)
2. Strong law of large numbers for any function f such that $|f(x)| \leq \pi^{-s}(x)$, $s \in (0,1)$. pioneering work by (Saksman & Vihola, 2009); we use many ideas of their paper!

Convergence of the (simplified) Equi-Energy sampler

We prove

- when the target density π is *lighter than exponential*, on a Polish space X
- whatever the nbr of stages, the probability of swap $\epsilon \in (0,1)$, the successive tempered distributions and the "hottest" one $\pi^{1/T_\star}$, $T_\star > 1$
- when the "first" auxiliary process is an ergodic Markov chain
- when P is a RWHM algorithm with Gaussian proposal distribution

1. Ergodicity: $\lim_n \mathbb{E}\left[f(X_n)\right] = \pi(f)$ for any bounded functions f.
2. Strong law of large numbers for any continuous function f such that $|f(x)| \leq \pi^{-s}(x)$, $s \in (0, 1/T_\star)$.

extensions of the works by (Atchadé, 2007), (Andrieu et al. 2009)

All the details in

G. Fort, E. Moulines, P. Priouret (2010). *Convergence of adaptive MCMC algorithms: ergodicity and law of large numbers*