

Stochastic approximation for adaptive Markov chain Monte Carlo algorithms

Gersende FORT

LTCI / CNRS - TELECOM ParisTech, France

I. Examples of adaptive and interacting MCMC samplers

1. Adaptive Hastings-Metropolis algorithm [HAARIO ET AL. 1999]
2. Equi-Energy algorithm [KOU ET AL. 2006]
3. Wang-Landau algorithm [WANG & LANDAU, 2001]

Adaptive Hastings-Metropolis algorithm

► Symmetric Random Walk Hastings-Metropolis algorithm

- Goal: sample a Markov chain with known stationary distribution π on \mathbb{R}^d (known up to a normalizing constant)
- Iterative mechanism: given the current sample X_n ,
 - propose a move to $X_n + Y$ $Y \sim q(\cdot - X_n)$
 - accept the move with probability

$$\alpha(X_n, X_n + Y) = 1 \wedge \frac{\pi(X_n + Y)}{\pi(X_n)}$$

and set $X_{n+1} = X_n + Y$; otherwise, $X_{n+1} = X_n$.

Adaptive Hastings-Metropolis algorithm

► Symmetric Random Walk Hastings-Metropolis algorithm

- Goal: sample a Markov chain with known stationary distribution π on \mathbb{R}^d (known up to a normalizing constant)
- Iterative mechanism: given the current sample X_n ,
 - propose a move to $X_n + Y$ $Y \sim q(\cdot - X_n)$
 - accept the move with probability

$$\alpha(X_n, X_n + Y) = 1 \wedge \frac{\pi(X_n + Y)}{\pi(X_n)}$$

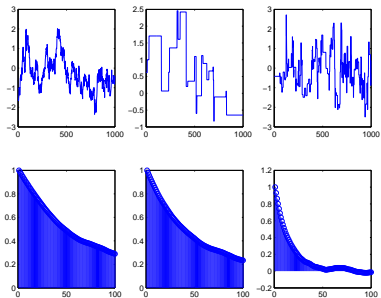
and set $X_{n+1} = X_n + Y$; otherwise, $X_{n+1} = X_n$.

- Design parameter: **how to choose the proposal distribution q ?**

For example, in the case $q(\cdot - x) = \mathcal{N}_d(x; \theta)$ how to scale the proposal i.e. how to choose the covariance matrix θ ?

“goldilock principle”

Too small, too large, better variance



► Adaptive Hastings-Metropolis algorithm(s)

Based on theoretical results [Gelman et al. 1996; · · ·] when the proposal is Gaussian $\mathcal{N}_d(x, \theta)$, choose θ

- as the covariance structure of π [Haario et al. 1999]: $\theta \propto \Sigma_\pi$. In practice, Σ_π is unknown and **this quantity is computed “online”** with the past samples of the chain

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

where μ_{n+1} is the empirical mean.

$\kappa > 0$, prevent from badly scaled matrix

► Adaptive Hastings-Metropolis algorithm(s)

Based on theoretical results [Gelman et al. 1996; . . .] when the proposal is Gaussian $\mathcal{N}_d(x, \theta)$, choose θ

- as the covariance structure of π [Haario et al. 1999]: $\theta \propto \Sigma_\pi$. In practice, Σ_π is unknown and **this quantity is computed “online”** with the past samples of the chain

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

where μ_{n+1} is the empirical mean. $\kappa > 0$, prevent from badly scaled matrix

- OR such that the mean acceptance rate converges to α_\star [Andrieu & Robert 2001]. In practice this θ is unknown and this parameter is **adapted during the run** of the algorithm

$$\theta_n = \tau_n \text{Id} \quad \text{with} \quad \log \tau_{n+1} = \log \tau_n + \gamma_{n+1} (\alpha_{n+1} - \alpha_\star)$$

where α_n is the mean acceptance rate.

- OR . . .

► In practice, **simultaneous adaptation** of the design parameter **and simulation**.

Given the current value of the chain X_n and the design parameter θ_n

- Draw the next sample X_{n+1} with the transition kernel $P_{\theta_n}(X_n, \cdot)$.
- Update the design parameter: $\theta_{n+1} = \Xi_{n+1}(\theta_n, X_{n+1}, \cdot)$.

- ▶ In practice, **simultaneous adaptation** of the design parameter **and simulation**.

Given the current value of the chain X_n and the design parameter θ_n

- Draw the next sample X_{n+1} with the transition kernel $P_{\theta_n}(X_n, \cdot)$.
- Update the design parameter: $\theta_{n+1} = \Xi_{n+1}(\theta_n, X_{n+1}, \cdot)$.

- ▶ In this MCMC context, we are interested in the behavior of the chain $\{X_n, n \geq 0\}$

e.g.

- Convergence of the marginals: $\mathbb{E}[f(X_n)] \rightarrow \pi(f)$ for f bounded.
- Law of large numbers: $n^{-1} \sum_{k=1}^n f(X_k) \rightarrow \pi(f)$ (a.s. or \mathbb{P})
- Central limit theorem

but

- we have $\pi P_\theta = \pi$ for any θ : all the transition kernels have the same inv. distribution π
- so, stability / convergence of the adaptation process $\{\theta_n, n \geq 0\}$ is not the main issue.

Equi-Energy sampler

- ▶ Proposed by Kou et al. (2006) for the simulation of multi-modal density π .

How to define a sampler that both allows

- local moves for a local exploration of the density.
- and large jumps in order to visit other modes of the target ?

Equi-Energy sampler

- ▶ Proposed by Kou et al. (2006) for the simulation of multi-modal density π .

How to define a sampler that both allows

- local moves for a local exploration of the density.
 - and large jumps in order to visit other modes of the target ?
- ▶ Idea: (a) build an **auxiliary process** that moves between the modes far more easily and (b) define the process of interest
- by running a “classical” MCMC algorithm
 - and sometimes, choose a value of the auxiliary process as the new value of the process of interest: draw a point at random + acceptance-rejection mechanism

Equi-Energy sampler

- ▶ Proposed by Kou et al. (2006) for the simulation of multi-modal density π .

How to define a sampler that both allows

- local moves for a local exploration of the density.
 - and large jumps in order to visit other modes of the target ?
- ▶ Idea: (a) build an **auxiliary process** that moves between the modes far more easily and (b) define the process of interest
- by running a “classical” MCMC algorithm
 - and sometimes, choose a value of the auxiliary process as the new value of the process of interest: draw a point at random + acceptance-rejection mechanism

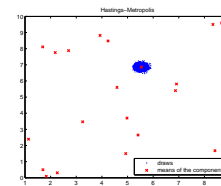
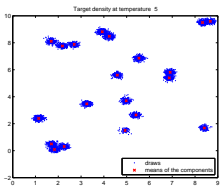
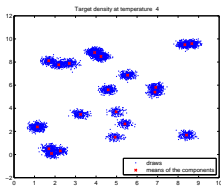
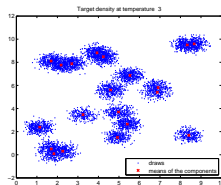
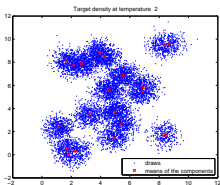
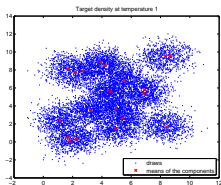
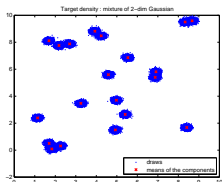
How to define such an auxiliary process ? Ans.: as a process with stationary distribution π^β ($\beta \in (0, 1)$), a **tempered** version of the target π .

► On an example: a K -stage Equi-Energy sampler.

• target density: $\pi = \sum_{i=1}^{20} \mathcal{N}_2(\mu_i, \Sigma_i)$

• K auxiliary processes: with targets π^{1/T_i}

$$T_1 > T_2 > \dots > T_{K+1} = 1$$



► Algorithm: (2 stages) Repeat:

- Update the adaptation process

$$\theta_n = \frac{1}{n} \sum_{k=0}^{n-1} \delta_{Y_k}$$

where $\{Y_n, n \geq 0\}$ is the auxiliary process with stationary distribution π^β .

- Update the process of interest with transition : $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$ where

$$P_{\theta_n}(x, A) = (1-\epsilon)P(x, A) + \epsilon \left\{ \int_A \underbrace{\alpha(x, y)}_{\text{accept/reject mechanism}} \theta_n(dy) + \delta_x(A) \int (1 - \alpha(x, y)) \theta_n(dy) \right\}$$

► Algorithm: (2 stages) Repeat:

- Update the adaptation process

$$\theta_n = \frac{1}{n} \sum_{k=0}^{n-1} \delta_{Y_k}$$

where $\{Y_n, n \geq 0\}$ is the auxiliary process with stationary distribution π^β .

- Update the process of interest with transition : $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$ where

$$P_{\theta_n}(x, A) = (1-\epsilon)P(x, A) + \epsilon \left\{ \int_A \underbrace{\alpha(x, y)}_{\text{accept/reject mechanism}} \theta_n(dy) + \delta_x(A) \int (1 - \alpha(x, y)) \theta_n(dy) \right\}$$

► In this example, $\pi P_\theta \neq \pi$ **BUT** $\pi P_{\pi^\beta} = \pi$ i.e. asymptotically, when θ_n "is" π^β , the process of interest $\{X_n, n \geq 0\}$ behaves like a Markov chain with invariant distribution π .

► Algorithm: (2 stages) Repeat:

- Update the adaptation process

$$\theta_n = \frac{1}{n} \sum_{k=0}^{n-1} \delta_{Y_k}$$

where $\{Y_n, n \geq 0\}$ is the auxiliary process with stationary distribution π^β .

- Update the process of interest with transition : $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$ where

$$P_{\theta_n}(x, A) = (1-\epsilon)P(x, A) + \epsilon \left\{ \int_A \underbrace{\alpha(x, y)}_{\text{accept/reject mechanism}} \theta_n(dy) + \delta_x(A) \int (1 - \alpha(x, y)) \theta_n(dy) \right\}$$

► In this example, $\pi P_\theta \neq \pi$ BUT $\pi P_{\pi^\beta} = \pi$ i.e. asymptotically, when θ_n "is" π^β , the process of interest $\{X_n, n \geq 0\}$ behaves like a Markov chain with invariant distribution π .

In this MCMC context, we are again interested in the behavior of $\{X_n, n \geq 0\}$ **but convergence of θ_n is crucial** since the algorithm is designed to "sample from π " only when $\theta_n = \pi^\beta$.

Wang-Landau algorithm

► Proposed by Wang & Landau (2001) to favor the moves between elements of a partition of the state space, when the weights of these elements are unknown.

- Goal:

- sample a chain on $\prod_{i=1}^d (X_i \times \{i\})$ with stationary distribution

$$\Pi(A_i \times \{i\}) = \frac{1}{d} \int_{A_i} \frac{h_i(x)}{\theta_*(i)} \mathbb{1}_{X_i}(x) dx ,$$

when θ_* is unknown

- and/ or estimate the normalizing constants $\theta_*(i)$.

Wang-Landau algorithm

► Proposed by Wang & Landau (2001) to favor the moves between elements of a partition of the state space, when the weights of these elements are unknown.

- Goal:

- sample a chain on $\prod_{i=1}^d (X_i \times \{i\})$ with stationary distribution

$$\Pi(A_i \times \{i\}) = \frac{1}{d} \int_{A_i} \frac{h_i(x)}{\theta_*(i)} \mathbb{1}_{X_i}(x) dx ,$$

when θ_* is unknown

- and/ or estimate the normalizing constants $\theta_*(i)$.

- Tool :

- A family of transition kernels P_θ on $\prod_{i=1}^d (X_i \times \{i\})$
 - where $\theta = (\theta(1), \dots, \theta(d))$ is a probability on $\{1, \dots, d\}$
 - with invariant distribution **known up to a normalizing constant**

$$\Pi_\theta(A_i \times \{i\}) = \left(\sum_{j=1}^d \frac{\theta_*(j)}{\theta(j)} \right)^{-1} \int_{A_i} \frac{\pi(x)}{\theta(i)} \mathbb{1}_{X_i}(x) dx ,$$

► Algorithm: repeat

- Draw $(X_{n+1}, I_{n+1}) \sim P_{\theta_n}((X_n, I_n), \cdot)$
- Update the adaptation process

$$\theta_{n+1}(i) \propto \theta_n(i) + \gamma_{n+1} \theta_n(i) \mathbb{1}_{I_{n+1}}(i)$$

► Algorithm: repeat

- Draw $(X_{n+1}, I_{n+1}) \sim P_{\theta_n}((X_n, I_n), \cdot)$
- Update the adaptation process

$$\theta_{n+1}(i) \propto \theta_n(i) + \gamma_{n+1} \theta_n(i) \mathbb{1}_{I_{n+1}}(i)$$

► In this MCMC context, we are also interested in the **convergence of the sequence** $\{\theta_n, n \geq 0\}$: at a first order,

$$\theta_{n+1}(i) \approx \theta_n(i) + \gamma_{n+1} \theta_n(i) \left(\mathbb{1}_{I_{n+1}}(i) - \theta_n(I_{n+1}) \right)$$

and when $(X_n, I_n) \sim \Pi_{\theta_n}$

$$\mathbb{E} \left[\theta_n(i) \left(\mathbb{1}_{I_{n+1}}(i) - \theta_n(I_{n+1}) \right) \mid \mathcal{F}_n \right] = \left(\sum_{j=1}^d \frac{\theta_*(j)}{\theta_n(j)} \right)^{-1} (\theta_*(i) - \theta_n(i))$$

i.e. $\{\theta_n, n \geq 0\}$ should converge to θ_* !

Conclusion (I)

In adaptive MCMC,

- given a family of transition kernels $\{P_\theta, \theta \in \Theta\}$
- ergodic with invariant distribution π_θ

we define a bivariate process $\{(X_n, \theta_n), n \geq 0\}$ such that

$$\mathbb{P}(X_{n+1} \in \cdot | \mathcal{F}_n) = P_{\theta_n}(X_n, \cdot)$$

θ_n is updated s.t. it *should converge* to θ_*

Two cases: $\pi_\theta = \pi$ for any θ OR $\pi_{\theta_*} = \pi$.

What kind of conditions on the adaptation mechanism, in order the process $\{X_n, n \geq 0\}$ to converge to the target distribution π ?

In the sequel, “convergence” means “convergence of the marginals”

$$\mathbb{E}[f(X_n)] \rightarrow \pi(f) \quad f \text{ bounded}$$

Conclusion (II)

Trois exemples illustrant des situations différentes :

① Hastings Metropolis adaptatif :

- tous les noyaux P_θ ont **même** mesure invariante π .

② Equi-Energy sampler :

- Chaque noyau P_θ a sa propre mesure invariante π_θ .
- On sait que π_θ existe mais on n'a pas d'expression explicite (régularité en $\theta \dots$)

③ Wang-Landau :

- Chaque noyau P_θ a sa propre mesure invariante π_θ .
- On a l'expression de π_θ (en fonction de θ).

II. Convergence of adaptive / interacting MCMC samplers

(Joint work with E. Moulines (Telecom ParisTech, France) and P. Priouret (Paris VI, France))

L'adaptation peut détruire la convergence

- Consider a family of transition kernels on $\{0, 1\}$:

$$P_\theta = \begin{pmatrix} 1 - \theta & \theta \\ \theta & 1 - \theta \end{pmatrix} \quad \theta \in (0, 1)$$

- Then, for any $\theta \in (0, 1)$, $\pi P_\theta = \pi$ with $\pi = (1/2; 1/2)$.
- Choose $t_0, t_1 \in (0, 1)$. Define the adaptive process:

$$\begin{cases} X_{n+1} \sim P_{\theta_n}(X_n, \cdot) \\ \theta_{n+1} = t_{X_{n+1}} \end{cases}$$

- Then, the transition kernel is $\begin{pmatrix} 1 - t_0 & t_0 \\ t_1 & 1 - t_1 \end{pmatrix}$
- and the invariant distribution is $\pi \propto (t_1, t_0)$.

Conditions pour la convergence

On écrit

$$\begin{aligned} \mathbb{E} [f(X_n)] - \pi(f) &= \mathbb{E} \left[f(X_n) - P_{\theta_{n-N}}^N f(X_{n-N}) \right] \\ &\quad + \mathbb{E} \left[P_{\theta_{n-N}}^N f(X_{n-N}) - \pi_{\theta_{n-N}}(f) \right] \\ &\quad + \mathbb{E} \left[\pi_{\theta_{n-N}}(f) \right] - \pi(f) \end{aligned}$$

Soit trois groupes de conditions

- ① Terme 1 : dû à l'adaptation: comparaison du processus adapté à une chaîne "gelée" (que l'on n'adapte plus).
- ② Terme 2 : ergodicité des noyaux de transitions P_{θ} .
- ③ Terme 3 : uniquement si $\pi_{\theta} \neq \pi$; c'est le plus délicat \dots surtout lorsque l'expression de π_{θ} n'est pas connue.

$$\begin{aligned} \mathbb{E}[f(X_n)] - \pi(f) &= \mathbb{E}\left[f(X_n) - P_{\theta_{n-N}}^N f(X_{n-N})\right] \\ &\quad + \mathbb{E}\left[P_{\theta_{n-N}}^N f(X_{n-N}) - \pi_{\theta_{n-N}}(f)\right] + \mathbb{E}\left[\pi_{\theta_{n-N}}(f)\right] - \pi_*(f) \end{aligned}$$

► [Terme 3] quand $\pi_\theta \neq \pi_*$, conditions pour $\lim_n \pi_{\theta_n}(f) = \pi_*(f)$

Puisque

$$\pi_{\theta_* + \Delta}(f) - \pi_{\theta_*}(f) = \pi_{\theta_*} (P_{\theta_* + \Delta} - P_{\theta_*}) (I - P_{\theta_*})^{-1} (I - \pi_{\theta_*})(f) + \text{“Reste”}$$

la convergence de $\{\pi_{\theta_n}(f), n \geq 0\}$ vers $\pi_{\theta_*}(f)$ est une conséquence de la convergence des noyaux P_{θ_n} vers P_{θ_*} .

$$\begin{aligned}\mathbb{E}[f(X_n)] - \pi(f) &= \mathbb{E}\left[f(X_n) - P_{\theta_{n-N}}^N f(X_{n-N})\right] \\ &\quad + \mathbb{E}\left[P_{\theta_{n-N}}^N f(X_{n-N}) - \pi_{\theta_{n-N}}(f)\right] + \mathbb{E}\left[\pi_{\theta_{n-N}}(f)\right] - \pi_*(f)\end{aligned}$$

► [Terme 3] quand $\pi_\theta \neq \pi_*$, conditions pour $\lim_n \pi_{\theta_n}(f) = \pi_*(f)$

Puisque

$$\pi_{\theta_*+\Delta}(f) - \pi_{\theta_*}(f) = \pi_{\theta_*}(P_{\theta_*+\Delta} - P_{\theta_*})(I - P_{\theta_*})^{-1}(I - \pi_{\theta_*})(f) + \text{“Reste”}$$

la convergence de $\{\pi_{\theta_n}(f), n \geq 0\}$ vers $\pi_{\theta_*}(f)$ est une conséquence de la convergence des noyaux P_{θ_n} vers P_{θ_*} .

- Cas favorable : convergence en “norme opérateur”.
- Sinon : résultats plus fins! Par exemple, si on sait que

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

que peut-on en déduire sur $\lim_n \pi_{\theta_n}(f)$?

Starting from :

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

Starting from :

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

the steps are:

$$\forall x \in X, \quad \exists \Omega_x, \quad \mathbb{P}(\Omega_x) = 1 \quad \forall \omega \in \Omega_x \quad \lim_n P_{\theta_n(\omega)}(x, \cdot) \xrightarrow{\mathcal{D}} P_{\theta_*}(x, \cdot)$$

↪ **Tool:** separable metric space X (ex. Polish)

Starting from :

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

the steps are:

$$\forall x \in X, \quad \exists \Omega_x, \quad \mathbb{P}(\Omega_x) = 1 \quad \forall \omega \in \Omega_x \quad \lim_n P_{\theta_n(\omega)}(x, \cdot) \xrightarrow{\mathcal{D}} P_{\theta_*}(x, \cdot)$$

↔ **Tool:** separable metric space X (ex. Polish)

$$\exists \Omega', \quad \mathbb{P}(\Omega') = 1 \quad \forall \omega \in \Omega', x \in X \quad \lim_n P_{\theta_n(\omega)}(x, \cdot) \xrightarrow{\mathcal{D}} P_{\theta_*}(x, \cdot) ,$$

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

Starting from :

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

the steps are:

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

↔ **Tool:** separable metric space X (ex. Polish)

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

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

$$\exists \Omega_\star, \quad \mathbb{P}(\Omega_\star) = 1 \quad \forall \omega \in \Omega_\star \quad \lim_n P_{\theta_n(\omega)}^k(x, \cdot) \xrightarrow{\mathcal{D}} P_{\theta_\star}^k(x, \cdot) ,$$

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

Starting from :

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

the steps are:

$$\forall x \in X, \quad \exists \Omega_x, \quad \mathbb{P}(\Omega_x) = 1 \quad \forall \omega \in \Omega_x \quad \lim_n P_{\theta_n(\omega)}(x, \cdot) \xrightarrow{\mathcal{D}} P_{\theta_*}(x, \cdot)$$

↪ **Tool:** separable metric space X (ex. Polish)

$$\exists \Omega', \quad \mathbb{P}(\Omega') = 1 \quad \forall \omega \in \Omega', x \in X \quad \lim_n P_{\theta_n(\omega)}(x, \cdot) \xrightarrow{\mathcal{D}} P_{\theta_*}(x, \cdot) ,$$

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

$$\exists \Omega_*, \quad \mathbb{P}(\Omega_*) = 1 \quad \forall \omega \in \Omega_* \quad \lim_n P_{\theta_n(\omega)}^k(x, \cdot) \xrightarrow{\mathcal{D}} P_{\theta_*}^k(x, \cdot) ,$$

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

Then

$$|\pi_{\theta_n}(f) - \pi_{\theta_*}(f)| \leq |P_{\theta_n}^k f(x) - \pi_{\theta_n}(f)| + |P_{\theta_*}^k f(x) - \pi_{\theta_*}(f)| + \left| P_{\theta_n}^k f(x) - P_{\theta_*}^k f(x) \right|$$

↪ **Tool:** ergodicity

$$\begin{aligned} \mathbb{E}[f(X_n)] - \pi(f) &= \mathbb{E}\left[f(X_n) - P_{\theta_{n-N}}^N f(X_{n-N})\right] \\ &\quad + \mathbb{E}\left[P_{\theta_{n-N}}^N f(X_{n-N}) - \pi_{\theta_{n-N}}(f)\right] + \mathbb{E}\left[\pi_{\theta_{n-N}}(f)\right] - \pi(f) \end{aligned}$$

► [Terme 2] condition on the ergodicity of the transition kernels “Usually”, the transition kernels $\{P_\theta, \theta \in \Theta\}$ are geometrically ergodic :

$$\sup_{f, |f|_\infty \leq 1} |P_\theta^n f(x) - \pi_\theta(f)| \leq C_\theta \rho_\theta^n V(x) \quad \rho_\theta \in (0, 1)$$

BUT the rate of convergence may depend upon $\theta \dots$ in such a way that

$$\rho_\theta \rightarrow 1 \quad \text{when } \theta \rightarrow \partial\Theta .$$

Therefore, the rate at which $\theta_n \rightarrow \partial\Theta$ has to be controlled.

En pratique,

- le contrôle de l'ergodicité se fait via des conditions de drift + minoration

Si

$$P_\theta V \leq \lambda_\theta V + b_\theta \quad P_\theta(x, \cdot) \geq \delta_\theta \nu_\theta(\cdot)$$

alors

$$\|P_\theta^n(x, \cdot) - \pi_\theta\|_{\text{TV}} \leq C_\theta \rho_\theta^n V(x)$$

où

$$C_\theta \vee (1 - \rho_\theta)^{-1} \leq C \left(b_\theta \vee \delta_\theta^{-1} \vee (1 - \lambda_\theta)^{-1} \right)^3$$

En pratique,

- le contrôle de l'ergodicité se fait via des conditions de drift + minoration

Si

$$P_\theta V \leq \lambda_\theta V + b_\theta \quad P_\theta(x, \cdot) \geq \delta_\theta \nu_\theta(\cdot)$$

alors

$$\|P_\theta^n(x, \cdot) - \pi_\theta\|_{\text{TV}} \leq C_\theta \rho_\theta^n V(x)$$

où

$$C_\theta \vee (1 - \rho_\theta)^{-1} \leq C \left(b_\theta \vee \delta_\theta^{-1} \vee (1 - \lambda_\theta)^{-1} \right)^3$$

- et la non-détérioration de ce contrôle quand $\theta \rightarrow \partial\Theta$ se règle au cas par cas
 - imposer que le paramètre θ "reste loin des bords" (reprojection dans un compact par ex.)
 - laisser vivre la procédure d'adaptation mais contrôler la croissance [Vihola & Saksman, 2010], [Vihola, 2010]

Ex. pour HM adaptatif, [Vihola & Saksman, 2010] montrent que

$$C_\theta \vee (1 - \rho_\theta)^{-1} \leq c\sqrt{\det\theta}$$

$$\forall \tau > 0, \quad n^{-\tau} |\theta_n| < +\infty \quad \text{p.s.}$$

$$\begin{aligned} \mathbb{E} [f(X_n)] - \pi(f) &= \mathbb{E} \left[f(X_n) - P_{\theta_{n-N}}^N f(X_{n-N}) \right] \\ &\quad + \mathbb{E} \left[P_{\theta_{n-N}}^N f(X_{n-N}) - \pi_{\theta_{n-N}}(f) \right] + \mathbb{E} \left[\pi_{\theta_{n-N}}(f) \right] - \pi(f) \end{aligned}$$

► [Terme 1] condition sur le mécanisme d'adaptation

since

$$\begin{aligned} &\left| \mathbb{E} \left[f(X_n) - P_{\theta_{n-N}}^N f(X_{n-N}) \right] \right| \\ &\leq \sum_{j=1}^{N-1} (N-j) \mathbb{E} \left[\underbrace{\sup_x \left\| P_{\theta_{n-N+j}}(x, \cdot) - P_{\theta_{n-N+j-1}}(x, \cdot) \right\|_{\text{TV}}}_{\text{"distance" between two successive transition kernels}} \right] \end{aligned}$$

Therefore, the adaptation has to be diminishing.

Adaptation and Ergodicity

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

► Example: $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$ $\theta_n = n^{-1/4}$ $P_\theta = \begin{pmatrix} 1 - \theta & \theta \\ \theta & 1 - \theta \end{pmatrix}$

In this case, since $\theta_n \rightarrow 0$

$$\left| \mathbb{E}\left[P_{\theta_{n-N}}^N f(X_{n-N}) - \pi(f)\right] \right| \leq |1 - 2\theta_{n-N}|^N \rightarrow 1 \quad N \text{ is fixed}$$

Adaptation and Ergodicity

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

► **Example:** $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$ $\theta_n = n^{-1/4}$ $P_\theta = \begin{pmatrix} 1 - \theta & \theta \\ \theta & 1 - \theta \end{pmatrix}$

In this case, since $\theta_n \rightarrow 0$,

$$\left| \mathbb{E}\left[P_{\theta_{n-N_n}}^{N_n} f(X_{n-N_n}) - \pi(f)\right] \right| \leq |1 - 2\theta_{n-N_n}|^{N_n} \rightarrow 0 \quad \text{for convenient } N_n$$

Therefore, we choose N depending upon n : $N_n \rightarrow +\infty$ and the adaptation has to be such that

$$\sum_{j=1}^{N_n-1} (N_n - j) \mathbb{E} \left[\underbrace{\sup_x \left\| P_{\theta_{n-N_n+j}}(x, \cdot) - P_{\theta_{n-N_n+j-1}}(x, \cdot) \right\|_{\text{TV}}}_{\text{"distance" between two successive transition kernels}} \right] \rightarrow 0$$

↔ The “rate” of adaptation depends on the ergodic behavior of the transition kernels

Stochastic approximation for adaptive Markov chain Monte Carlo algorithms

└ Convergence of adaptive/interacting MCMC samplers

└ Adaptation and Ergodicity

III. Conclusion

Tools for convergence of adaptive MCMC samplers

- 1 Markov chain theory (ergodicity, Poisson equation, \dots)
- 2 Stochastic approximation (stability/convergence, control of “non-stability”)

Tools for convergence of adaptive MCMC samplers

- 1 Markov chain theory (ergodicity, Poisson equation, \dots)
 - 2 Stochastic approximation (stability/convergence, control of “non-stability”)
- When the transition kernels have the same invariant distribution π
- Ergodicity of transition kernels.
 - Diminishing adaptation

Ex. convergence of $\{\theta_n, n \geq 0\}$ is not required BUT the control of “divergence to $\partial\Theta$ ” is needed

Tools for convergence of adaptive MCMC samplers

- 1 Markov chain theory (ergodicity, Poisson equation, \dots)
 - 2 Stochastic approximation (stability/convergence, control of “non-stability”)
- When the transition kernels have the same invariant distribution π
- Ergodicity of transition kernels.
 - Diminishing adaptation
- Ex. convergence of $\{\theta_n, n \geq 0\}$ is not required BUT the control of “divergence to $\partial\Theta$ ” is needed
- When they have their own invariant distribution and $\pi_{\theta_\star} = \pi$.
- Ergodicity, Diminishing adaptation
 - Convergence of θ_n to θ_\star

Procédures d'approximation stochastique

Est-il nécessaire de modifier l'adaptation pour que la procédure d'approximation stochastique

- soit récurrente
- soit p.s. bornée (stabilité)
- converge vers l'ensemble d'intérêt.

par exemple en introduisant

- une reprojction sur un compact fixe
- une reprojction sur des compacts croissants
- ...

doublée d'une "troncation" de la chaîne

↪ pas toujours utile de forcer la récurrence / stabilité puisqu'on sait s'accomoder d'une non-stabilité du paramètre ...

↪ travaux de recherche en cours pour éviter ces reprojctions / troncations.

Some references

1 Adaptive MCMC (methodologies)

- C. Andrieu, J. Thoms. *Statistics and Computing*, 2008.
- J.S. Rosenthal. *MCMC handbook*, 2009.
- Y. Atchadé, G. Fort, E. Moulines, P. Priouret. *Time sSeries book*, 2010.

2 General results for convergence of adaptive MCMC

- C. Andrieu, E. Moulines. *Ann. Appl. Probab.*, 2006.
- G.O. Roberts, J.S. Rosenthal. *J. Appl. Probab.* 2007.
- Y. Atchadé, G. Fort. *Stoch. Processes Appl.*, 2010.
- G. Fort, E. Moulines, P. Priouret. *Preprint*, 2010.

3 Convergence of some adaptive MCMC

- C. Andrieu, A. Jasra, A. Doucet, P. DelMoral *Preprint 2007*
- Y. Atchadé *Statistica Sinica*, 2010
- E. Saksman, M. Vihola, *Ann. Appl. Probab.* 2010
- M. Vihola *Preprint*, 2010