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 mecanism: given the current sample X_n,
  - propose a move to $X_n + Y$ $Y \sim q(\cdot - X_n)$ where $q(-z) = q(z)$
  - 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 mecanism: given the current sample X_n,
    - propose a move to $X_n + Y$ $\quad Y \sim q(\cdot - X_n)$ $\quad$ where $q(-z) = q(z)$
    - accept the move with probability

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

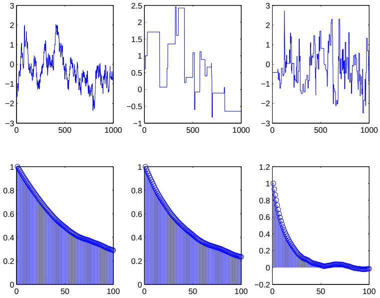
    and set $X_{n+1} = X_n + Y$; otherwise, $X_{n+1} = X_n$.

  - 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

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

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; $\cdots$] 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

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

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

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

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

  where α_n is the mean acceptance rate.

- OR $\cdots$

▶ 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: $\mathbb{E}\left[f(X_n)\right] \longrightarrow \pi(f)$ for f bounded.
- Law of large numbers: $n^{-1} \sum_{k=1}^{n} f(X_k) \longrightarrow \pi(f)$ (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 + acceptation-rejection mecanism

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

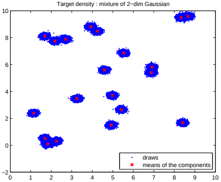
- local moves for a local exploration of the density.
- and large jumps in order to visit other modes of the target ?

▶ 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 + acceptation-rejection mecanism

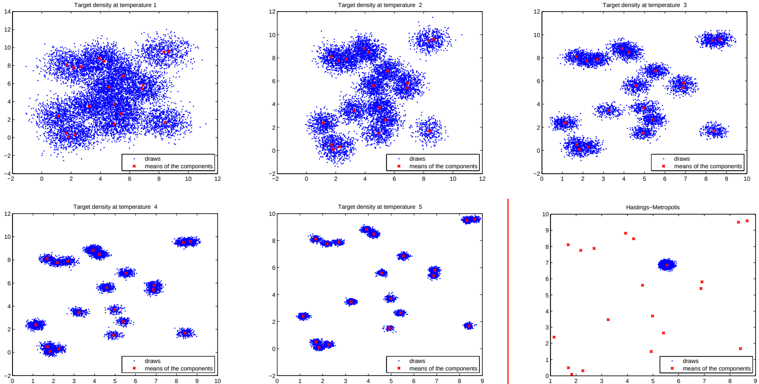
How to define such an auxiliary process ? Ans.: as a process with stationary distribution π^β ($\beta \in (0,1)$), a tempered version of the target π.

- 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}

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



- 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

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

- 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

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

- 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

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

- 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 $\prod_{i=1}^{d}(\mathsf{X}_i \times \{i\})$ with stationary distribution
    $$\Pi(A_i \times \{i\}) = \frac{1}{d} \int_{A_i} \frac{h_i(x)}{\theta_\star(i)} \mathbb{1}_{\mathsf{X}_i}(x)\ dx \ ,$$
    when $\theta_\star$ is unknown
  - and/ or estimate the normalizing constants $\theta_\star(i)$.

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 $\prod_{i=1}^{d} (\mathsf{X}_i \times \{i\})$ with stationary distribution
$$\Pi(A_i \times \{i\}) = \frac{1}{d} \int_{A_i} \frac{h_i(x)}{\theta_\star(i)} \mathbb{1}_{\mathsf{X}_i}(x)\ dx \ ,$$
    when $\theta_\star$ is unknown
  - and/ or estimate the normalizing constants $\theta_\star(i)$.

- Tool :
  - A family of transition kernels P_θ on $\prod_{i=1}^{d} (\mathsf{X}_i \times \{i\})$
  - where $\theta = (\theta(1), \cdots, \theta(d))$ is a probability on $\{1, \cdots, d\}$
  - with invariant distribution known up to a normalizing constant
$$\Pi_\theta(A_i \times \{i\}) = \left(\sum_{j=1}^{d} \frac{\theta_\star(j)}{\theta(j)} \right)^{-1} \int_{A_i} \frac{\pi(x)}{\theta(i)} \mathbb{1}_{\mathsf{X}_i}(x)\ dx \ ,$$

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

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

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

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

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

θ_n is updated s.t. it *should converge* to $\theta_\star$

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

What kind of conditions on the adaptation mecanism, 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"

$$\mathbb{E}\left[f(X_n)\right] \to \pi(f) \qquad f \text{ bounded}$$

Conclusion (II)

Trois exemples illustrant des situations différente :

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 $\theta \cdots$)

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))

Adaptation can be really bad for convergence

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

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

  - Then, for any $\theta \in (0,1)$, $\quad \pi P_\theta = \pi$ 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 for the convergence

We write

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

Three sets of conditions:

1. Term 1: is null when no adaptation. Comparison of the adapted process to a "frozen" chain (i.e. a chain for which we stop adaptation).
2. Term 2: ergodicity of the transition kernels P_θ.
3. Term 3: only if $\pi_\theta \neq \pi$; it is the most difficult (i.e. technical!) step $\cdots$ mainly in the case π_θ isnot known in closed form.

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

- [Term 3] when $\pi_\theta \neq \pi_\star$, conditions so that $\lim_n \pi_{\theta_n}(f) = \pi_\star(f)$

Since

$$\pi_{\theta_\star+\Delta}(f) - \pi_{\theta_\star}(f) = \pi_{\theta_\star}\left(P_{\theta_\star+\Delta} - P_{\theta_\star}\right)\left(I - P_{\theta_\star}\right)^{-1}\left(I - \pi_{\theta_\star}\right)(f) + \text{"Remainder term"}$$

the convergence of $\{\pi_{\theta_n}(f), n \geq 0\}$ to $\pi_{\theta_\star}(f)$ is a consequence of the convergence of the transition kernels P_{θ_n} to $P_{\theta_\star}$.

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

▶ [Term 3] when $\pi_\theta \neq \pi_\star$, conditions so that $\lim_n \pi_{\theta_n}(f) = \pi_\star(f)$

Since

$$\pi_{\theta_\star+\Delta}(f) - \pi_{\theta_\star}(f) = \pi_{\theta_\star}\left(P_{\theta_\star+\Delta} - P_{\theta_\star}\right)\left(I - P_{\theta_\star}\right)^{-1}\left(I - \pi_{\theta_\star}\right)(f) + \text{"Remainder term"}$$

the convergence of $\{\pi_{\theta_n}(f), n \geq 0\}$ to $\pi_{\theta_\star}(f)$ is a consequence of the convergence of the transition kernels P_{θ_n} to $P_{\theta_\star}$.

- Easiest situation: convergence in "operator norm".
- Otherwise: quite technical results! for example, when we know

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

what can be said on the convergence of $\lim_n \pi_{\theta_n}(f)$?

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:

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

$\hookrightarrow$ 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:

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

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

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

$\hookrightarrow$ 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:

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

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

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

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

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

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

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:

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

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

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

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

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

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

Then

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

$\hookrightarrow$ Tool: ergodicity

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

▶ [Term 2] condition on the ergodicity of the transition kernels "Usually", the transition kernels $\{P_\theta, \theta \in \Theta\}$ are geometrically ergodic :

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

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

$$\rho_\theta \longrightarrow 1 \qquad \text{when} \quad \theta \to \partial\Theta \ .$$

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

In practice,

- control of ergodicity is a consequence of drift conditions and minorization conditions

  If
  $$P_\theta V \leq \lambda_\theta V + b_\theta \qquad\qquad P_\theta(x,\cdot) \geq \delta_\theta \nu_\theta(\cdot)$$
  then
  $$\|P_\theta^n(x,\cdot) - \pi_\theta\|_{\mathrm{TV}} \leq C_\theta \ \rho_\theta^n \ V(x)$$
  where
  $$C_\theta \vee (1-\rho_\theta)^{-1} \leq C \left(b_\theta \vee \delta_\theta^{-1} \vee (1-\lambda_\theta)^{-1} \right)^3$$

In practice,

- control of ergodicity is a consequence of drift conditions and minorization conditions

  If

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

  then

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

  where

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

- and, in order to prove that this control does not degenerate when $\theta \to \partial\Theta$, quite ad-hoc conditions
  - do such that the "parameter" θ "remains far from the boundaries" (reprojection on compact sets for instance)
  - OR do not modify the procedure but, control the rate at which the parameter tends to the boundaries [Vihola & Saksman, 2010], [Vihola, 2010]

  Ex. for adaptive HM, [Vihola & Saksman, 2010] show that

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

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

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

- [Term 1] condition on the adaptation mecanism

since

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

Therefore, the adaptation has to be diminishing.

Adaptation and Ergodicity

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

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

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

$$\left|\mathbb{E}\left[P^N_{\theta_{n-N}} f(X_{n-N}) - \pi(f)\right]\right| \leq |1 - 2\theta_{n-N}|^N \to 1 \qquad N \text{ is fixed}$$

Adaptation and Ergodicity

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

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

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

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

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

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

$\hookrightarrow$ The “rate” of adaptation depends on the ergodic behavior of the transition kernels

III. Conclusion

Tools for convergence of adaptive MCMC samplers

1. Markov chain theory (ergodicity, Poisson equation, $\cdots$)
2. Stochastic approximation (stability/convergence, control of "non-stability")

Tools for convergence of adaptive MCMC samplers

1. Markov chain theory (ergodicity, Poisson equation, $\cdots$)
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 $\partial\Theta$" in needed

Tools for convergence of adaptive MCMC samplers

1. Markov chain theory (ergodicity, Poisson equation, $\cdots$)
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 $\partial\Theta$" in needed

- When they have their own invariant distribution and $\pi_{\theta_\star} = \pi$.
  - Ergodicity, Diminishing adaptation
  - Convergence of θ_n to $\theta_\star$

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

doublée d'une "troncation" de la chaîne

$\hookrightarrow$ pas toujours utile de forcer la récurrence / stabilité puisqu'on sait s'accomoder d'une non-stabilité du paramètre $\cdots$

$\hookrightarrow$ travaux de recherche en cours pour éviter ces reprojections / 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*