Adaptive and Interacting MCMC algorithms

Eric MOULINES

Telecom Paris Tech
CNRS - LTCI

Joint work with G. FORT (TELECOM ParisTech, France), P. PRIOURET (Univ. Paris VI, France), P. VANDEKHERKOVE (Univ. Marne la Vallée)

Outline

1. MCMC algorithms are a flexible family of algorithms to sample distributions, known up to a normalisation factor,
2. This flexibility comes at a price... badly tuned MCMC can be very slow to converge and the convergence may be difficult to diagnose.
3. In the last 10 years, several classes of algorithms have been introduced to *increase* the sampling efficiency of the MCMC, without demanding much additional user supervision. The common idea is to let the algorithms **self-learned** from the past simulations by **adapting** its parameters
4. **Problem :** the Markov property is not retained and the convergence is more difficult to study
5. **Today :** the basic ingredients of successful adaptations.

Adaptive MCMC

The Dream : Given a model (i.e., X and π), the computer :

- efficiently and cleverly tries out different MCMC algorithms ;
- automatically **learns** the good ones ;
- runs the algorithm for **long enough** ;
- obtains excellent estimates together with error bounds ;
- reports the results clearly and concisely, while user unaware of the complicated MCMC and adaption that was used.

The Reality : Easier said than done !

Adaptive MCMC

- Let $\{P_\theta, \theta \in \Theta\}$ be a collection of Markov chain kernels on X, each of which is ϕ-irreducible and aperiodic and has $\pi(\cdot)$ as a stationary distribution :

$$\pi P_\theta = \pi \ , \quad \text{for any } \theta \in \Theta$$

- The parameter space Θ the **parameter space** can either be **finite dimensional** or **infinite dimensional**.
- Let θ_n be a sequence of Θ-valued random variables which are updated according to specific rules.
- **Assumption** the adaptation is **conditionally** Markovian, *i.e.* θ_n **resume** all the informations obtained in the past to adapt the proposal

$$\mathbb{P}[X_{n+1} \in A | \mathcal{G}_n] = P_{\theta_n}(X_n, A)$$

where $\mathcal{G}_n = \sigma(X_0, \ldots, X_n, \theta_0, \ldots, \theta_n)$.

An elementary example : the Adaptive Metropolis Algorithm

- $Y_{k+1} = X_k + Z_{k+1}$ where $Z_{k+1} \sim_{\text{i.i.d.}} \bar{q}$, and q is symmetric, $\bar{q}(z) = \bar{q}(-z)$
- In this case, $q(x, y) = q(y, x) = \bar{q}(y - x) = \bar{q}(x - y)$ and the acceptance rate does not depend on the proposal distribution

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

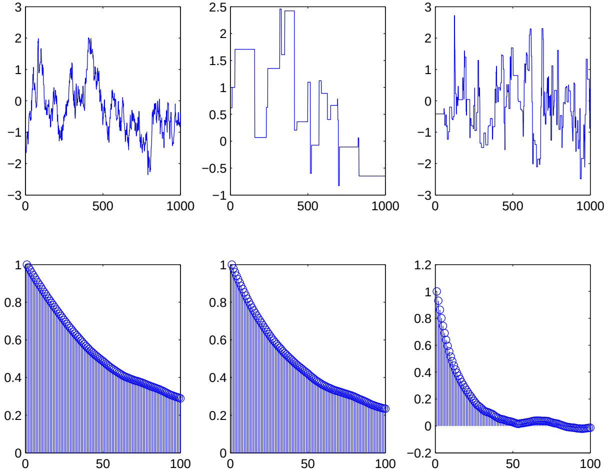
- ... biased random walk where some moves get rejected.

Influence of the scaling

- If the variance is either **too small** or **too large**, then the convergence rate of the Markov chain will be slow and any inference from values drawn from the chain are likely to be unreliable.
  1. **too small**... almost all the proposal are accepted. Nevertheless, the stepsizes are small, and the algorithm visits the state space very slowly.
  2. **too large**... many propositions fall in regions where π is very small. These proposals are often rejected and the algorithm get stuck at a point.

Finding a proper scale is thus mandatory ! but it is not always obvious to say what **small** or **large** mean for a given distribution π and a given function.

Scaling


3
2
1
0
−1
−2
−3
0
500
1000
2.5
2
1.5
1
0.5
0
−0.5
−1
0
500
1000
3
2
1
0
−1
−2
−3
0
500
1000
1
0.8
0.6
0.4
0.2
0
0
50
100
1
0.8
0.6
0.4
0.2
0
0
50
100
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
0
50
100

Optimization Criterion I

- To find a proper **scaling**, a criterion reflecting the performance of the sampling algorithm is required.
- In practical MCMC, interest may lie in the estimation of a (certain numbers) of additive functionals. For any one of these functionals, f say, a plausible criterion to minimise is the stationary integrated autocorrelation time for f under π

$$\tau(f) = 1 + 2\sum_{i=1}^{\infty} \mathrm{Corr}_\pi(f(X_0), f(X_i))$$

- The central limit theorem for additive functional of Markov chains $\{f(X_i)\}$ gives a Monte Carlo variance proportional to $\tau(f)$...

Optimization Criterion II

- This approach has two major disadvantages...
  1. Estimation of $\tau(f)$ is notoriously difficult... This is equivalent to estimating the spectral density of the process $\{f(X_k)\}$ at zero frequency at stationarity...
  2. The optimisation criterion gives a different solution for the **optimal** chain for different functionals f.
- Several approaches have been proposed to obtain a more easy to compute and more generally meaningful optimisation criterion...
- The common idea is to replace the optimisation problem by a simpler one, which captures the salient features of the original problem.

Optimal Scaling of the RWM

- A useful idea is to consider a **high-dimensional** limit... By rescaling the time as a function of the dimension, a diffusion limit can be obtained.
- The choice of an optimal scaling then translates into the optimization of the speed this limiting diffusion.
- In the diffusion limit, the problem of non-uniqueness of the optimum is avoided since in the limit the correlation $\tau(f)$ is proportional to the inverse of the speed of the limiting diffusion...

Diffusive Limits

- Stationary distribution : $\pi^{(d)}(x_1, \ldots, x_d) = \prod_{i=1}^{d} f(x_i)$ on $\mathbb{R}^d$ (asymptotic $= d \to \infty$)
- Metropolis proposal : $q_\theta^{(d)}(x_1, \ldots, x_d) \sim \mathcal{N}\left(0, (\theta^2/d)\mathrm{I}_d\right)$... with variance decreasing as $1/d$.
- Interpolated process : $Z_t^{(d)} = X_{[td],1}^{(d)}$... we consider a single component and we speed up the time scale by d.
- When d becomes large, a single component basically see the mean of the others (**mean-field**)...

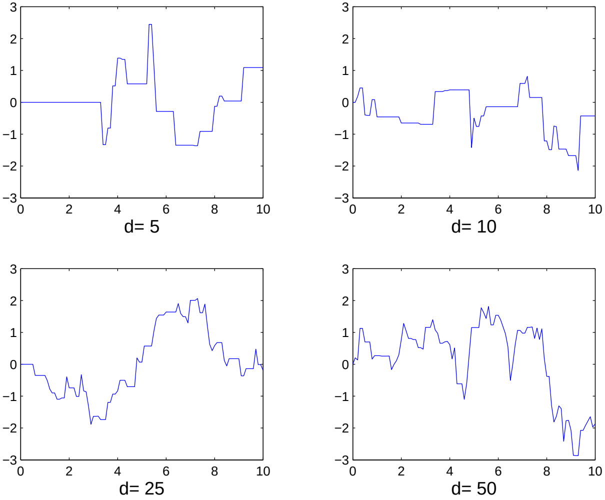


FIGURE: Diffusive limits for different values of d

Diffusive Limits

- $Z^{(d)} \Rightarrow Z$, where Z solves the Langevin SDE

$$dZ_t = v^{1/2}(\theta)dB_t + (1/2)v(\theta)\nabla \log f(Z_t)dt$$
$$v(\theta) = 2\theta^2 \Phi\left(-\theta\sqrt{I}/2\right)$$

  where Φ is the distribution function of $\mathcal{N}(0,1)$ and I is Fisher Information of the translation model associated to f, $I = \int (f'(x)/f(x))^2 f(x)dx$.

- $v(\theta)$ is the speed of the diffusion : $Z_t = \tilde{Z}_{v(\theta)t}$ where $\{\tilde{Z}_t\}$ is a solution of the normalized Langevin SDE

$$d\tilde{Z}_t = dB_t + (1/2)\nabla \log f(\tilde{Z}_t)dt.$$

Speed / Acceptance rate

- Mean Acceptance rate (stationary regime)

$$\tau^{(d)}(\theta) = \iint \pi^{(d)}(\mathbf{x}) q_\theta^{(d)}(\mathbf{y}-\mathbf{x}) \left\{ 1 \wedge \frac{\pi^{(d)}(\mathbf{y})}{\pi^{(d)}(\mathbf{x})} \right\} d\mathbf{x} d\mathbf{y} \ .$$

- Result : $\tau^{(\infty)}(\theta) = \lim_{d \to \infty} \tau^{(d)}(\theta)$ exists and it is possible to relate the speed of the diffusion to the mean acceptance rate !

$$v(\theta) = \tau^{(\infty)}(\theta) \left\{ \Phi^{-1}(\tau^{(\infty)}(\theta)/2) \right\}^2$$

- The speed is optimal for the value θ_* of the parameter which satisfies $\tau^{(\infty)}(\theta_*) = \bar{\tau} \approx 0.234$...

Pros and Cons of diffusion limits

- Empirically this 0.234 rule has been observed to be approximately right much more generally.
- Extensions and generalisations of this result can be found in (Roberts and Rosenthal, 2001) and (Bedard, 2007), (Pillai, Stuart, 2009), (Bedard, Douc, Fort, Moulines, 2010).
- The focus of much of this work is in trying to characterise when the 0.234 rule holds and to explain how and why it breaks down in other situations.
- One major disadvantage of the diffusion limit work is its reliance on asymptotics in the dimensionality of the problem. Although it is often empirically observed that the limiting behaviour can be seen in rather small dimensional problems, (see for example Gelman et al., 1996), it is difficult to quantify this in any general way.

How to control the Acceptance Rate

- Objective : Finding the scaling factor θ solving

$$h(\theta) \overset{\mathrm{def}}{=} \iint \alpha(x,y) q_\theta(y-x) \pi(x) \mathrm{d}x \mathrm{d}y - \bar{\tau} = 0,$$

  where $\alpha(x,y) = \{1 \wedge \pi(y)/\pi(x)\}$.
- Under general assumptions, $\theta \to h(\theta)$ is monotone with $\lim_{\theta \to 0^+} h(\theta) = 1 - \bar{\tau} > 0$ and $\lim_{\theta \to \infty} h(\theta) = -\bar{\tau} < 0$... But $h(\theta)$ cannot be computed explicitly !
- Nevertheless, denoting θ_k the scaling value at iteration k, $\alpha(X_k, Y_{k+1}) - \bar{\tau}$ may be seen as a **noisy** observation of $h(\theta_k)$...
- **Suggest to use a stochastic approximation procedure to tune θ.**

Controlled Metropolis Algorithm

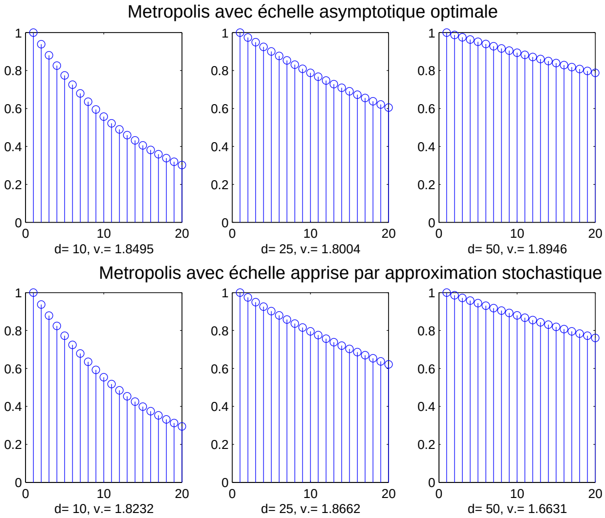
- Proposition & Accept/Reject

$$Y_{k+1} = X_k + \theta_k \mathcal{N}(0, \mathrm{Id})$$
$$X_{k+1} = \begin{cases} Y_{k+1} & \text{with prob. } \alpha(X_k, Y_{k+1}) \\ X_k & \text{otherwise} \end{cases}$$

- Update the scaling factor

$$\log(\theta_{k+1}) = \log(\theta_k) + \gamma_{k+1} \left\{ \alpha(X_k, Y_{k+1}) - \bar{\tau} \right\}$$

where $\lim_{k\to\infty} \gamma_k = 0$ and $\sum_{k=1}^{\infty} \gamma_k = \infty$.



Metropolis avec échelle asymptotique optimale

d= 10, v.= 1.8495

d= 25, v.= 1.8004

d= 50, v.= 1.8946

Metropolis avec échelle apprise par approximation stochastique

d= 10, v.= 1.8232

d= 25, v.= 1.8662

d= 50, v.= 1.6631

Multidimensional scaling

- Same asymptotic analysis ($d \to \infty$) with

$$\pi^{(d)}_{\Sigma_d}(\mathbf{x}) = |\Sigma_d|^{-1} \pi^{(d)}\left(\Sigma_d^{-1}\mathbf{x}\right), \quad \pi^{(d)}(x_1, \ldots, x_d) = \prod_{i=1}^{d} f(x_i)$$

$$q \sim N(0, (\sigma^2/d)\mathrm{Id})$$

then $Z_t^{(d)} = X_{[td],1}$ converges to the solution a Langevin SDE.

- the target acceptance rate (0.234...) which maximizes the speed of the limiting diffusion is independent from Σ_d, but the achievable maximal speed is strongly affected by Σ_d... loss

$$\lim_d \frac{d^{-1} \sum_{i=1}^d \lambda_{d,i}^2}{\left(d^{-1} \sum_{i=1}^d \lambda_{d,i}\right)^2}$$

where $\lambda_{d,i}$ eigenvalues of Σ_d.

Adaptive MCMC with multidim. scaling

1. Simulate

$$Y_{k+1} = X_k + \mathcal{N}(0, \sigma_k \Gamma_{\mathrm{k}})$$

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

2. Update the target mean and covariance

$$\mu_{k+1} = \mu_k + \gamma_{k+1}(X_{k+1} - \mu_k)$$

$$\Gamma_{k+1} = \Gamma_k + \gamma_{k+1}\left\{(X_{k+1} - \mu_k)(X_{k+1} - \mu_k)^T - \Gamma_k\right\}$$

3. Control the global scale of the proposal

$$\sigma_{k+1} = \sigma_k + \gamma_{k+1}\left(\alpha(X_k, Y_{k+1}) - \bar{\tau}\right)$$

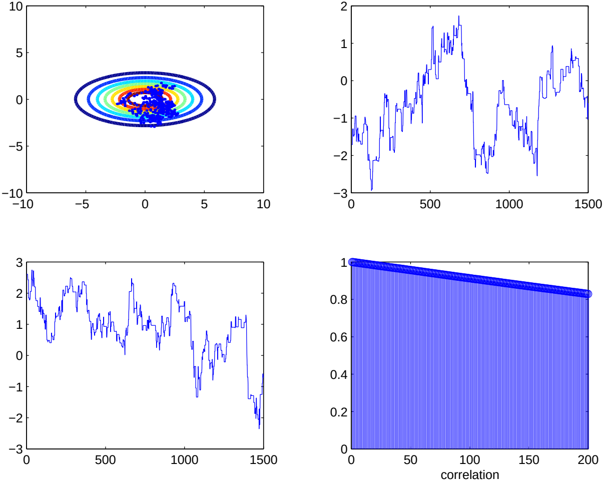


FIGURE: $d = 12$, $\pi \sim \mathcal{N}(0, \Gamma)$, $\mathrm{cond}(\Gamma) \approx 100$, $q \sim \mathcal{N}(0, (2.32^2/d)\,\mathrm{I})$

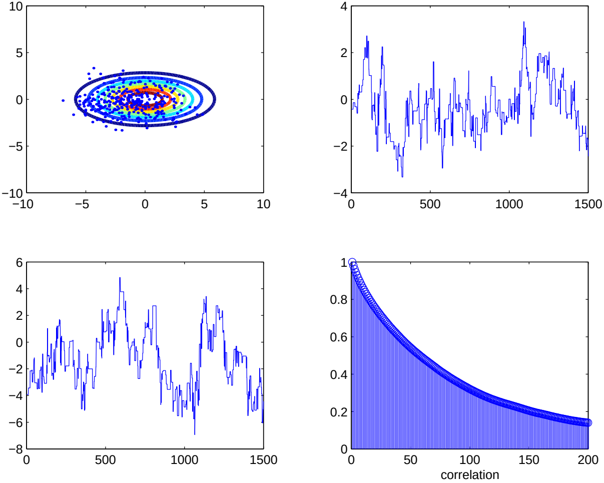


FIGURE: $d = 12$, $\pi \sim \mathcal{N}(0, \Gamma)$, $\mathrm{cond}(\Gamma) \approx 100$, $q \sim \mathcal{N}(0, (2.32^2/d)\,\Gamma)$

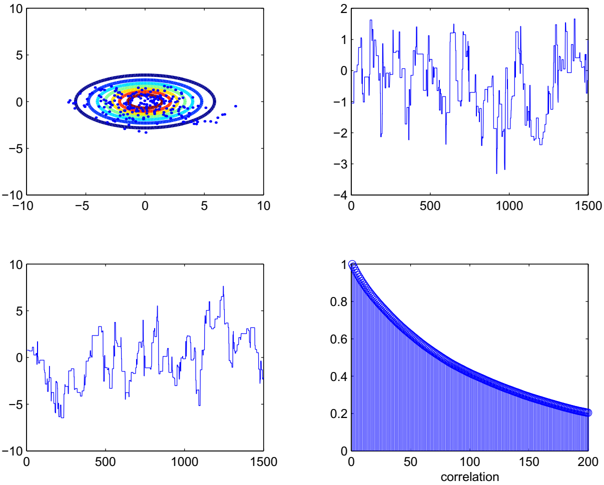


FIGURE: $d = 12$, $\pi \sim \mathcal{N}(0, \Gamma)$, $\mathrm{cond}(\Gamma) \approx 100$, $q \sim \mathcal{N}(0, \sigma_k \Gamma_k)$, with adaptive multidimensional scaling

Adaptive MCMC

- A family of transition kernels $\{P_\theta, \theta \in \Theta\}$ such that, for all $\theta \in \Theta$, the target distribution $\pi_\star$ is the stationary distribution of P_θ : $\pi_\star P_\theta = \pi_\star$.
- An adaptive MCMC algorithm : process $\{(X_n, \theta_n), n \geq 0\}$ on the product space $\mathsf{X} \times \Theta$:
  - **Sampling** : given the past, draw
    $$X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$$
  - **Internal adaptation** : update the **parameter** θ_n from the **past** values of the X and θ

Interacting MCMC

- a transition kernel P s.t. $\pi_\star P = \pi_\star$
- a probability of swap $\epsilon \in (0,1)$
- an auxiliary process $\{Y_n, n \geq 0\}$ targeting a **tempered** version $\pi_\star^\beta$
- Iteration n :

(a) with probability $(1-\epsilon)$ draw $X_{n+1} \sim P(X_n, \cdot)$

$$P_{\theta_n}(X_n, A) = (1-\epsilon) P(X_n, A) + \cdots$$

Interacting MCMC

- a transition kernel P s.t. $\pi_\star P = \pi_\star$
- a probability of swap $\epsilon \in (0,1)$
- an auxiliary process $\{Y_n, n \geq 0\}$ targeting a **tempered** version $\pi_\star^\beta$
- Iteration n :

(b) with probability ϵ, **draw** a point $Y_\star$ among $\{Y_1, \cdots, Y_n\}$ and **accept/reject** with probability $\alpha(X_n, Y_\star)$

$$P_{\theta_n}(X_n, A) = (1-\epsilon) P(X_n, A) + \epsilon \left\{ \int_A \theta_n(dy)\ \alpha(X_n, y) \right. \\ \left. + \mathbb{1}_A(X_n) \int \theta_n(dy)\ \{1 - \alpha(X_n, y)\} \right\}$$

where

$$\theta_n(dy) = \frac{1}{n} \sum_{k=1}^n \delta_{Y_k}(dy) \quad \text{and} \quad \alpha(x,y) = 1 \wedge \frac{\pi(y)\ \theta_\star(x)}{\theta_\star(y)\ \pi(x)}$$

An example of application

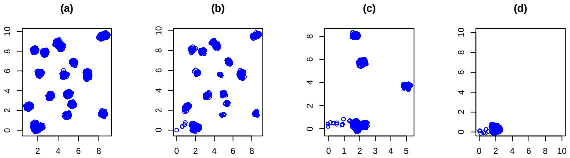


FIGURE: Example : Mixture of a 2D-Normal distribution [target / EE / Parallel Tempering / SRWM]

Interacting MCMC

- Construct an auxiliary process $\{Y_n, n \geq 0\}$ s.t. its empirical process $\lim_n \theta_n$ converges in some appropriate sense to $\theta_\star(\cdot)$ so that asymptotically,
$$P_{\theta_n} \approx P_{\theta_\star}$$
- The acceptance ratio $\alpha(x, y)$ of the interaction is chosen s.t. $\pi_\star \; P_{\theta_\star} = \pi_\star$
- **Heuristic** :
  1. if these two conditions are satisfied, then the distribution of $(X_k)_{k \geq 0}$ converges to $\pi_\star$ as $k \to \infty$.
  2. **wishful thinking** : The interaction speed up the convergence... the gain in convergence speed is large enough to offset the cost of sampling from an auxiliary process.

Interacting MCMC : refinements

When sampling from the past of the auxiliary process, select the points : introduce a **selection** $g(x,y)$ function (satisfying $g(x,y)=g(y,x)$)

$$P_{\theta_n}(X_n,A) = (1-\epsilon)P(X_n,A) + \epsilon \left\{ \int_A \frac{g(x,y)\theta_n(dy)}{\int g(x,y)\theta_n(dy)} \alpha(X_n,y) \right. \\ \left. + \mathbb{1}_A(X_n) \int \frac{g(x,y)\theta_n(dy)}{\int g(x,y)\theta_n(dy)} \{1-\alpha(X_n,y)\} \right\}$$

where

$$\theta_n(dy) = \frac{1}{n}\sum_{k=1}^{n} \delta_{Y_k}(dy) \qquad \alpha(x,y) = 1 \wedge \frac{\pi(y)\ \theta_\star(x)}{\tilde{\pi}(y)\ \pi(x)}$$

Interacting MCMC : refinements

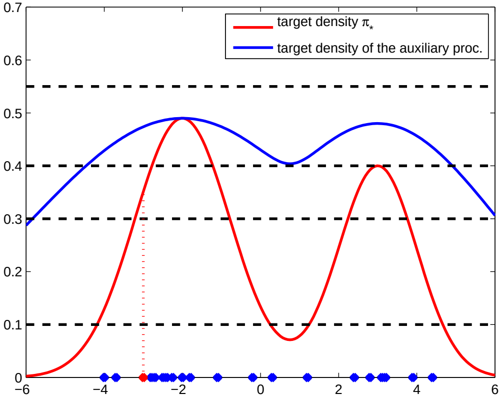
When sampling from the past of the auxiliary process, select the points : introduce a **selection** $g(x,y)$ function (satisfying $g(x,y) = g(y,x)$)

$$P_{\theta_n}(X_n, A) = (1 - \epsilon_{\theta_n}(x)) P(X_n, A) + \epsilon_{\theta_n}(x) \left\{ \int_A \frac{g(x,y)\theta_n(dy)}{\int g(x,y)\theta_n(dy)} \alpha(X_n, y) \right.$$
$$\left. + \mathbb{1}_A(X_n) \int \frac{g(x,y)\theta_n(dy)}{\int g(x,y)\theta_n(dy)} \{1 - \alpha(X_n, y)\} \right\}$$

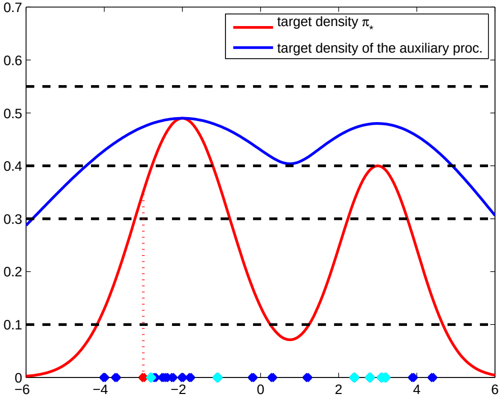
where

$$\theta_n(dy) = \frac{1}{n} \sum_{k=1}^{n} \delta_{Y_k}(dy) \;, \alpha(x,y) = 1 \wedge \frac{\pi(y)\ \theta_\star(x)}{\theta_\star(y)\ \pi(x)} \;, \epsilon_\theta(x) = \epsilon \mathbb{1}_{\int \theta(dy) g(x,y) > 0}.$$

The equi-energy sampler



The equi-energy sampler



Regularized Interacting MCMC

- Instead on drawing from θ_n, this distribution can be regularized by using a kernel

$$\tilde{\theta}_n(x) = (nh_n^d)^{-1} \sum_{i=1}^{n} K\left(\frac{x - Y_i}{h_n}\right) ,$$

  where d is the dimension of the space, (h_n) is a sequence of positive numbers and K is a Borel measurable function (kernel) satisfying $K \geq 0$ and $\int K = 1$.
- Provided that $\lim_{n\to\infty} h_n = 0$ and $\lim_{n\to\infty} nh_n^d = \infty$, $J_n \overset{\mathrm{def}}{=} \int |\tilde{\theta}_n(x) - \theta_\star(x)| \mathrm{d}x \to 0$ as $n \to \infty$ (for all $\epsilon > 0$, there exists n_0 s.t. for all $n \geq n_0$, $\mathbb{P}(J_n \geq \epsilon) \leq \mathrm{e}^{-rn}$).
- If K is the Gaussian kernel, then drawing from $\tilde{\theta}_n$ instead of θ_n amounts to add to add an independent Gaussian random variable with covariance $h_n^{1/2}\mathrm{Id}$... which is almost for free !

Interacting MCMC

- A family of transition kernels $\{P_\theta, \theta \in \Theta\}$ with invariant probability distribution π_θ : $\pi_\theta P_\theta = \pi_\theta$
- An interacting MCMC is a process $\{(X_n, \theta_n), n \geq 0\}$ on the product space $\mathsf{X} \times \Theta$ defined as
  - **Simulation** Given the past , draw
$$X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$$
  - **External adaptation** update the **parameter** θ_n (here, a probability distribution) according to
$$\theta_{n+1} \longleftrightarrow \text{computed from } \underline{\text{an auxiliary}} \text{ process } \{Y_k, k \leq n\}$$

Adaptive and interacting MCMC in a nutshell

- A family of transition kernels $\{P_\theta, \theta \in \Theta\}$ with invariant distribution : $\pi_\star$ (**internal adaptation**) or π_θ (**external adaptation**).

We define a filtration $\mathcal{F}_n$, and a process $\{(X_n, \theta_n), n \geq 0\}$ s.t.

- component θ_n : $\mathcal{F}_n$ adapted with **internal / external** adaptation
- component X_n (process of interest) :

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

Convergence of the marginals

- **Key ingredients** to prove the ergodicity of an MCMC algorithms :
  1. Markov Chain
  2. the transition kernel is **reversible** w.r.t the target distribution
- These properties are lost when adapting the algorithms...
- **Questions :** Conditions to guarantee that the **adaptation does not destroy the convergence** ?

Adaptive MCMC : $\pi_\theta = \pi_\star$

$$
\begin{aligned}
\mathbb{E}\left[f(X_n)\right] &= \mathbb{E}\left[\mathbb{E}\left[f(X_n)|\mathcal{F}_{n-N}\right]\right] \\
&= \mathbb{E}\Bigg[\underbrace{\mathbb{E}\left[f(X_n)|\mathcal{F}_{n-N}\right] - P^N_{\theta_{n-N}} f(X_{n-N})}_{\text{comparison with a frozen chain with transition } P_{\theta_{n-N}}} \\
&\qquad + \underbrace{P^N_{\theta_{n-N}} f(X_{n-N}) - \pi_\star(f)}_{\text{ergodicity of the frozen chain}}\Bigg] + \pi_\star(f).
\end{aligned}
$$

Diminishing adaptation

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

- Generally problem specific
- ... But most often, amounts to check a condition of the type

$$\sup_x \|P_{\theta_n}(x,\cdot) - P_{\theta_{n-1}}(x,\cdot)\|_{\mathrm{TV}} \leq C \;\|\theta_n - \theta_{n-1}\|_{\mathbf{xxx}}$$

so that convergence in probability is implied by the **adaptation scheme**.

Containment condition

$$\lim_M \limsup_n \mathbb{P}\left(M_\epsilon(X_n, \theta_n) \geq M\right) = 0,$$

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

- Most often, deduced from **ergodicity** + homogeneity
- The easy case is when the ergodicity is **uniform** in θ :

$$\sup_\theta \|P_\theta^n(x,\cdot) - \pi_\star\|_{\mathrm{TV}} \leq \rho(n)\ U(x) \qquad\qquad \lim_n \rho(n) = 0$$

then

$$M_\epsilon(x,\theta) \leq \rho^{-1}\left(\epsilon C^{-1} U^{-1}(x)\right).$$

Adaptive MCMC

Theorem

Assume

1. **(Diminishing adaptation)**

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

2. **(Containment condition)**

$$\lim_M \limsup_n \mathbb{P}\left(M_\epsilon(X_n,\theta_n) \geq M\right) = 0.$$

Then

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

Interacting MCMC : $\pi_\theta P_\theta = \pi_\theta$

$$
\begin{aligned}
\mathbb{E}\left[f(X_n)\right] &= \mathbb{E}\left[\mathbb{E}\left[f(X_n)|\mathcal{F}_{n-N}\right]\right] \\
&= \mathbb{E}\Bigg[\underbrace{\mathbb{E}\left[f(X_n)|\mathcal{F}_{n-N}\right] - P^N_{\theta_{n-N}} f(X_{n-N})}_{\text{comparison with a frozen chain with transition } P_{\theta_{n-N}}} \\
&\quad + \underbrace{P^N_{\theta_{n-N}} f(X_{n-N}) - \pi_{\theta_{n-N}}(f)}_{\text{ergodicity of the frozen chain}} \\
&\quad + \pi_{\theta_{n-N}}(f) - \pi_\star(f)\Bigg] + \pi_\star(f).
\end{aligned}
$$

Interacting MCMC : $\pi_\theta P_\theta = \pi_\theta$

$$\begin{aligned}
\mathbb{E}\left[f(X_n)\right] &= \mathbb{E}\left[\mathbb{E}\left[f(X_n)|\mathcal{F}_{n-N}\right]\right] \\
&= \mathbb{E}\Bigg[\underbrace{\mathbb{E}\left[f(X_n)|\mathcal{F}_{n-N}\right] - P^N_{\theta_{n-N}} f(X_{n-N})}_{\text{comparison with a frozen chain with transition } P_{\theta_{n-N}}} \\
&\quad + \underbrace{P^N_{\theta_{n-N}} f(X_{n-N}) - \pi_{\theta_{n-N}}(f)}_{\text{ergodicity of the frozen chain}} \\
&\quad + \pi_{\theta_{n-N}}(f) - \pi_\star(f)\Bigg] + \pi_\star(f).
\end{aligned}$$

- (same) : Diminishing adaptation, Containment condition
- **Convergence of the invariant measures** $\{\pi_{\theta_n}, n \geq 0\}$ to some $\pi_\star$

Interacting MCMC

Theorem

Assume

1. *(Diminishing adaptation)*

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

2. **(Containment condition)**

$$\lim_M \limsup_n \mathbb{P}\left(M_\epsilon(X_n,\theta_n) \geq M\right) = 0.$$

3. *(Convergence of the invariant distributions)*

$$\pi_{\theta_n}(f) - \pi_\star(f) \to_{\mathbb{P}} 0.$$

Then

$$\lim_n \; |\mathbb{E}\left[f(X_n)\right] - \pi_\star(f)| = 0$$

Spectral theory of V-uniformly ergodic operators

- Consider that X is a topological space endowed with its Borel σ-field. For any positive function V, denote by

$$\mathcal{C}_V = \left\{ f \text{ continuous}, \|f\|_V = \sup_{x \in \mathsf{X}} \frac{|f(x)|}{V(x)} < \infty \right\} .$$

- Denote by $\|P\|_V = \sup_{\|f\|_V \leq 1} \|Pf\|_V$ the operator norm.

Spectral theory of V-uniformly ergodic operators

- **Assumption** : P is a Feller transition kernel on X which is ψ-irreducible and aperiodic. Furthermore, there exist positive numbers $\epsilon > 0$ and C and a measurable function V, bounded on compact sets and unbounded out of compact sets, s.t. $PV(x) \leq (1-\epsilon)V(x) + b\mathbb{1}_C(x)$.
- P is a Markov transition probability with a unique invariant probability π such that, for some function $V \geq 1$, $\|P\|_V < \infty$ and the spectral radius of $P - \pi$ in $\mathcal{C}_V$ is smaller than one : $\|P^n - \pi\|_V \leq C\rho^n$.

Continuity of the spectrum

- Since the eigenvalue 1 is separated from the rest of the spectrum (with multiplicity 1), this part of the spectrum changes with P continuously, just as in the finite dimensional space
- Denote by $\Sigma(P)$ the spectrum and $\Sigma'(P) = \Sigma(P) \setminus \{1\}$. $\Sigma(P)$ can be separated from 1 by a closed curve Γ.
- The operator P can be decomposed as $P = \Pi + R$, where the spectrum $\Sigma(R)$ (on $\mathcal{C}_V$) is outside the domain enclosed by Γ, $\Pi R = R\Pi = 0$ and Π is a rank one operator

$$\Pi f(x) = \pi(f) \mathbb{1}(x) \; .$$

Continuity of the spectrum

- Assume that (P_n) is a sequence of Markov transition converging to P in the operator norm.
- Then, for n sufficiently large, $P_n = \Pi_n + R_n$, the spectrum of R_n (on $\mathcal{C}_V$) is outside the domain enclosed by Γ, $\Pi_n R_n = R_n \Pi_n = 0$ where Π_n is a rank one operator

$$\Pi_n f(x) = \pi_n(f) \mathbb{1}(x) \; .$$

- In addition,

$$\Pi_n = \frac{1}{2\mathrm{i}\pi} \int_\Gamma (\lambda - P_n)^{-1} \mathrm{d}\lambda$$

  and the condition $\| P_n - P \|_V \to 0$ implies that $\| \pi_n - \pi \|_V \to 0$...

Continuity of the spectrum

- According to the discussion above, the condition $\|P_{\theta_n} - P_{\theta_\star}\|_V \longrightarrow 0$ implies that $\|\pi_{\theta_n} - \pi_{\theta_\star}\|_V \longrightarrow 0$.
- This is enough for the **regularized** version of the interacting MCMC... but this is not enough for the original version of the equi-energy sampler
- A little bit more is needed to analyse the original equi-energy sampler

Back to the Interacting MCMC

Let $\pi_\star$ be positive and continuous on X s.t. $\sup_{\mathsf{X}} \pi_\star < +\infty$ and let $\beta \in (0,1)$.

- **On the auxiliary process :**
- **On the transition kernel P :**

Back to the Interacting MCMC

Let $\pi_\star$ be positive and continuous on X s.t. $\sup_{\mathsf{X}} \pi_\star < +\infty$ and let $\beta \in (0,1)$.

- **On the auxiliary process :** for any bounded function f,

$$\frac{1}{n}\sum_{k=1}^{n} f(Y_k) \longrightarrow_{a.s.} \pi_\star^\beta(f).$$

- **On the transition kernel P :**

Back to the Interacting MCMC

Let $\pi_\star$ be positive and continuous on X s.t. $\sup_{\mathsf{X}} \pi_\star < +\infty$ and let $\beta \in (0,1)$.

- **On the auxiliary process :** for any bounded function f,

$$\frac{1}{n} \sum_{k=1}^{n} f(Y_k) \longrightarrow_{a.s.} \pi_\star^{\beta}(f).$$

- **On the transition kernel P :** P is phi-irreducible, $\pi_\star P = \pi_\star$, the level sets $\{\pi \geq p\}$ are 1-small and

$$PV(x) \leq \lambda V(x) + b \mathbb{1}_{\mathcal{C}}(x) \qquad V(x) = \left(\frac{\pi(x)}{\sup_{\mathsf{X}} \pi} \right)^{-\tau(1-\beta)}$$

for some $\lambda \in (0,1)$, $b < +\infty$, a set $\mathcal{C}$, $\tau \in (0,1]$.

Under these conditions,

- the **diminishing adaptation** condition holds
- the **containment condition** holds.
- the invariant measures a.s. converge : $\lim_n \pi_{\theta_n}(f) = \pi_\star(f)$ a.s. for any bounded Lipshitz function.

Hence, for any bounded function f

$$\mathbb{E}\left[f(X_n)\right] \longrightarrow_n \pi_\star(f).$$

Strong LLN

Sufficient Conditions for the existence of $\pi_\star$ s.t. the strong LLN

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \longrightarrow_{a.s.} \pi_\star(f)$$

is satisfied for any function f in a (hopefully large) class of functions.

Idea : use the Poisson equation

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) - \pi_\star(f) = \underbrace{\frac{1}{n}\sum_{k=1}^{n}\{f(X_k) - \pi_{\theta_{k-1}}(f)\}}_{\text{“Poisson term”}}$$

$$+ \quad \underbrace{\frac{1}{n}\sum_{k=1}^{n}\pi_{\theta_{k-1}}(f) - \pi_\star(f)}_{\text{Cesaro mean (is null when } \pi_\theta = \pi_\star)}$$

The first step consists in proving that

$$\pi_{\theta_n}(f) \longrightarrow_{a.s.} \pi_{\theta_\star}(f) \qquad \text{for any } f \in \mathcal{L}_{V^\alpha},\ \alpha \in [0,1)$$

Decomposition

$$
\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n}\{f(X_k) - \pi_{\theta_{k-1}}(f)\} \\
&\quad = n^{-1}\underbrace{\sum_{k=1}^{n}\{\hat{f}_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}}\hat{f}_{\theta_{k-1}}(X_{k-1})\}}_{\text{martingale term}} \\
&\quad + \underbrace{\frac{1}{n}\sum_{k=1}^{n-1}\{P_{\theta_k}\hat{f}_{\theta_k}(X_k) - P_{\theta_{k-1}}\hat{f}_{\theta_{k-1}}(X_k)\}}_{\text{Remainder term (I)}} \\
&\quad + \underbrace{n^{-1}\{P_{\theta_0}f_{\theta_0}(X_0) - P_{\theta_{n-1}}f_{\theta_{n-1}}(X_{n-1})\}}_{\text{Remainder term (II)}}
\end{aligned}
$$

where $\hat{f}_\theta$ solves $$f - \pi_\theta(f) = \hat{f}_\theta - P_\theta\hat{f}_\theta.$$

Decomposition

$$
\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n}\{f(X_k)-\pi_{\theta_{k-1}}(f)\} \\
&\quad = n^{-1}\underbrace{\sum_{k=1}^{n}\{\hat{f}_{\theta_{k-1}}(X_k)-P_{\theta_{k-1}}\hat{f}_{\theta_{k-1}}(X_{k-1})\}}_{\text{martingale term}} \\
&\quad + \underbrace{\frac{1}{n}\sum_{k=1}^{n-1}\{P_{\theta_k}\hat{f}_{\theta_k}(X_k)-P_{\theta_{k-1}}\hat{f}_{\theta_{k-1}}(X_k)\}}_{\text{Remainder term (I)}} \\
&\quad + \underbrace{n^{-1}\{P_{\theta_0}f_{\theta_0}(X_0)-P_{\theta_{n-1}}f_{\theta_{n-1}}(X_{n-1})\}}_{\text{Remainder term (II)}}
\end{aligned}
$$

where $\hat{f}_\theta$ solves $\qquad f-\pi_\theta(f)=\hat{f}_\theta-P_\theta\hat{f}_\theta.$

- a.s. convergence of the martingale : conditions on the L^p-moments of the increment $\hookrightarrow$ **uniform-in-θ drift conditions** on the

Decomposition

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

$$= n^{-1}\underbrace{\sum_{k=1}^{n}\{\hat{f}_{\theta_{k-1}}(X_k)-P_{\theta_{k-1}}\hat{f}_{\theta_{k-1}}(X_{k-1})\}}_{\text{martingale term}}$$

$$+\underbrace{\frac{1}{n}\sum_{k=1}^{n-1}\{P_{\theta_k}\hat{f}_{\theta_k}(X_k)-P_{\theta_{k-1}}\hat{f}_{\theta_{k-1}}(X_k)\}}_{\text{Remainder term (I)}}$$

$$+\underbrace{n^{-1}\{P_{\theta_0}f_{\theta_0}(X_0)-P_{\theta_{n-1}}f_{\theta_{n-1}}(X_{n-1})\}}_{\text{Remainder term (II)}}$$

where $\hat{f}_\theta$ solves $\qquad f-\pi_\theta(f)=\hat{f}_\theta-P_\theta\hat{f}_\theta.$

- a.s. convergence of the martingale : conditions on the L^p-moments of the increment $\hookrightarrow$ **uniform-in-θ drift conditions** on the

Define

$$D_V(\theta, \theta') := \sup_x \frac{\|P_\theta(x, \cdot) - P_{\theta'}(x, \cdot)\|_V}{V(x)}$$

Theorem

Assume

(i) **(uniform ergodic behavior)** P_θ *is phi-irreducible,*

$$P_\theta V \leq \lambda V + b\mathbb{1}_{\mathcal{C}} \qquad \lambda \in (0,1), b < +\infty,$$

and level sets of V *are* 1*-small.*

(ii) *(strengthened D.A.)* $\sum_k \frac{1}{k} V^\alpha(X_k) \; D_{V^\alpha}(\theta_k, \theta_{k-1}) < +\infty$ *a.s.*

(iii) *(convergence of the invariant measures)*

Then : if $\mathbb{E}[V(X_0)] < \infty$*, for any* $\alpha \in [0,1)$ *and any* $f \in \mathcal{L}_{V^\alpha}$

$$\frac{1}{n} \sum_{k=1}^n f(X_k) \longrightarrow_{a.s.} \pi_\star(f),$$

Conclusion

- We prove convergence of the marginals for adaptive and interacting MCMC samplers with the main ingredients
  - diminishing adaptation
  - ergodicity of the kernels + some form of uniformity in θ
  - For external adaptation : a.s. convergence of the invariant measures π_{θ_n}
- Under the same assumptions, a L.L.N can be established.