Adaptive and Interacting Markov chain Monte Carlo

Gersende FORT

LTCI
CNRS & Telecom ParisTech
Paris, France

Talk based on joint works with

Eric Moulines (Telecom ParisTech, France), Pierre Priouret (Univ. Paris 6, France) and Pierre Vandekerkhove (Univ. Marne-la-Vallée, France).

Amandine Schreck (Telecom ParisTech, France).

Benjamin Jourdain (ENPC, France), Estelle Kuhn (INRA, France), Tony Lelièvre (ENPC, France) and Gabriel Stoltz (ENPC, France).

Hastings-Metropolis algorithm (1/2)

Given

- a target density π on $\mathbb{X} \subseteq \mathbb{R}^d$ (to simplify the talk)
- a proposal transition kernel $q(x, y)$

define $\{X_k, k \geq 0\}$ iteratively as

(i) draw $Y \sim q(X_k, \cdot)$

(ii) compute

$$\alpha(X_k, Y) = 1 \wedge \frac{\pi(Y)}{\pi(X_k)} \frac{q(Y, X_k)}{q(X_k, Y)}$$

(iii) set $X_{k+1} = \begin{cases} Y & \text{with prob. } \alpha(X_k, Y) \\ X_k & \text{with prob. } 1 - \alpha(X_k, Y) \end{cases}$

Hastings-Metropolis algorithm (2/2)

Then $(X_k)_{k\geq 0}$ is a Markov chain with transition kernel P

$$P(x,A) = \int \alpha(x,y) q(x,y) \lambda(dy) + \mathbb{1}_A(x) \int (1-\alpha(x,y))\, q(x,y)\, \lambda(dy)$$

Under conditions on π and q

- Ergodic behavior : $P^k(x,\cdot) \xrightarrow{d} \pi$
- Explicit control of ergodicity $\|P^k(x,\cdot) - \pi\|_{\mathrm{TV}} \leq B(x,k)$
- Law of Large Numbers

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

- Central Limit Theorem

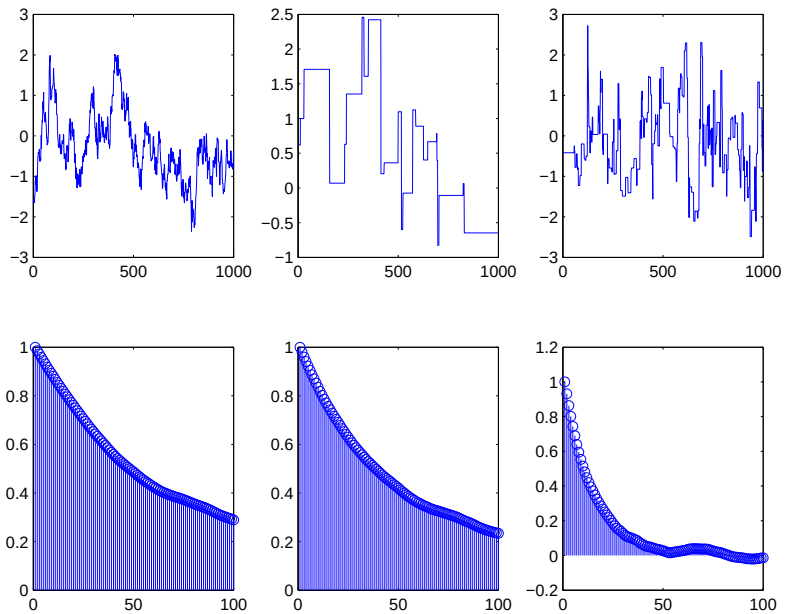
$$\sqrt{n}\left(\frac{1}{n}\sum_{k=1}^{n} f(X_k) - \int f\, \pi\, d\lambda\right) \xrightarrow{d} \mathcal{N}(0, \sigma_f^2)$$

ex. : Efficiency of a Gaussian Random Walk Hastings-Metropolis

When $\lambda \equiv$ Lebesgue on $\mathbb{R}$ and $\quad q(x, \cdot) \equiv \mathcal{N}(x, \theta)$

efficiency compared through the (estimated) lag-s autocovariance function

$$\gamma_s = \mathbb{E}\left[X_0 X_s\right] - \left(\mathbb{E}\left[X_0\right]\right)^2 \qquad \text{when} X_0 \sim \pi$$



For 3 different values of θ : [top] a path $(X_k, k \geq 1)$ [bottom] $s \mapsto \gamma(s)/\gamma(0)$

$\hookrightarrow$ **Online Adaption** of the design parameters θ

Example 1 : Adaptive Metropolis (1/2)

- Proposed by Haario et al. (2001) : learn on the fly the optimal covariance of the Gaussian proposal distribution

- Define a process $\{X_k, k \geq 0\}$ such that

(i) update the chain :

$\mathbb{P}\left(X_{k+1} \in A | \mathcal{F}_k\right) \equiv$ one step of Gaussian HM, with covariance matrix θ_k

(ii) update the estimate of the covariance matrix

$$\theta_{k+1} = \text{function}\left(k, \theta_k, X_{k+1}\right).$$

Example 1 : Adaptive Metropolis (2/2)

The general framework :

- Let P_θ be a Gaussian Hastings-Metropolis kernel ; θ is the covariance matrix of the Gaussian proposal distribution.
- For any θ : $\pi P_\theta = \pi$

The adaptive algorithm :

(i) Sample

$$X_{k+1}|\mathcal{F}_k \sim P_{\theta_k}(X_k, \cdot)$$

(ii) Update the parameter θ_{k+1} by using θ_k, X_{k+1}.

Here, θ is a covariance matrix.

Example 2 : Wang-Landau (1/4)

- Proposed by Wang and Landau (2001) for sampling systems in molecular dynamics ; many metastable states $\leftrightarrow$ many local modes separated with deep valleys.

- Idea : Let $\mathbb{X}_1, \cdots, \mathbb{X}_d$ be a partition of $\mathbb{X}$. Set

$$\pi_{\theta_\star}(x) \propto \sum_{i=1}^{d} \frac{\pi(x)}{\theta_\star(i)} \mathbb{1}_{\mathbb{X}_i}(x) \qquad \theta_\star(i) = \pi(\mathbb{X}_i)$$

  The idea is to obtain samples (approx.) under $\pi_{\theta_\star}$. Then, by an importance ratio, these samples will approximate π.

$$\text{roughly :} \qquad \frac{1}{n}\sum_{k=1}^{n} \delta_{X_k} \approx \pi_{\theta_\star} \Longrightarrow \frac{1}{n}\sum_{k=1}^{n} \theta_\star(i) \mathbb{1}_{X_k \in \mathbb{X}_i}\ \delta_{X_k} \approx \pi$$

- WL is an algorithm which provides an estimation of $\theta_\star$ and samples approx. distributed under $\pi_{\theta_\star}$.

Example 2 : Wang-Landau (2/4)

- Define $\{X_k, k \geq 0\}$ iteratively

  (i) Sample

  $$X_{k+1}|\mathcal{F}_k \sim \text{MCMC sampler with target distribution } \pi_{\theta_k}$$

  (ii) Update the parameter

  $$\theta_{k+1} = \text{function}\,(k, \theta_k, X_{k+1})$$

- The parameter $\{\theta_k, k \geq 0\}$ is updated through a Stochastic Approximation procedure $\theta_{n+1} = \theta_n + \gamma_{n+1} h(\theta_n) + \gamma_{n+1}\,\text{noise}_{n+1}$ with mean field h such that if $\{\theta_k, k \geq 0\}$ converges, its limiting value is $\theta_\star$.

Example 2 : Wang-Landau (3/4)

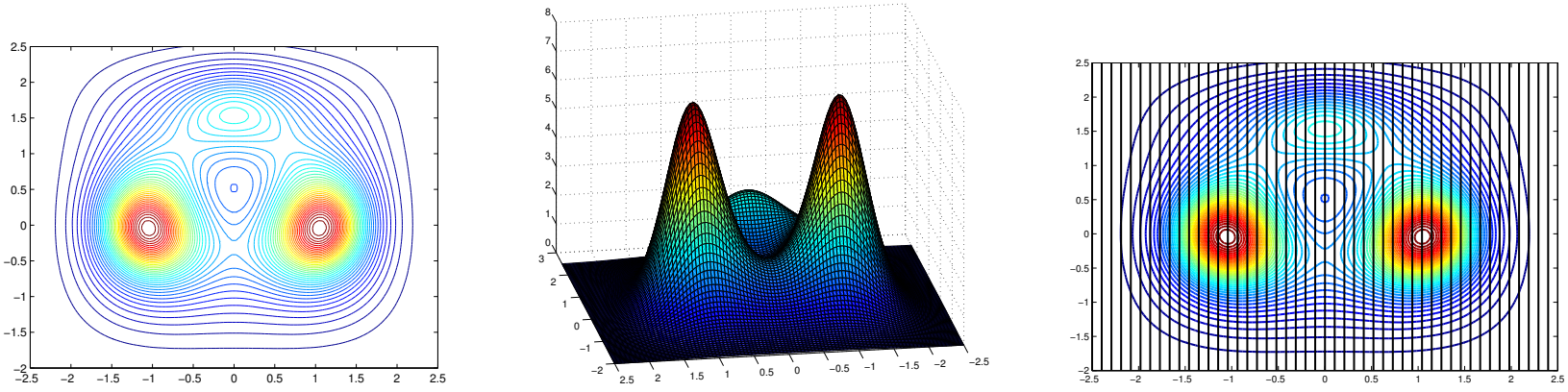


FIGURE: [left] level curves of π [center] Target density π [right] Partition of the state space

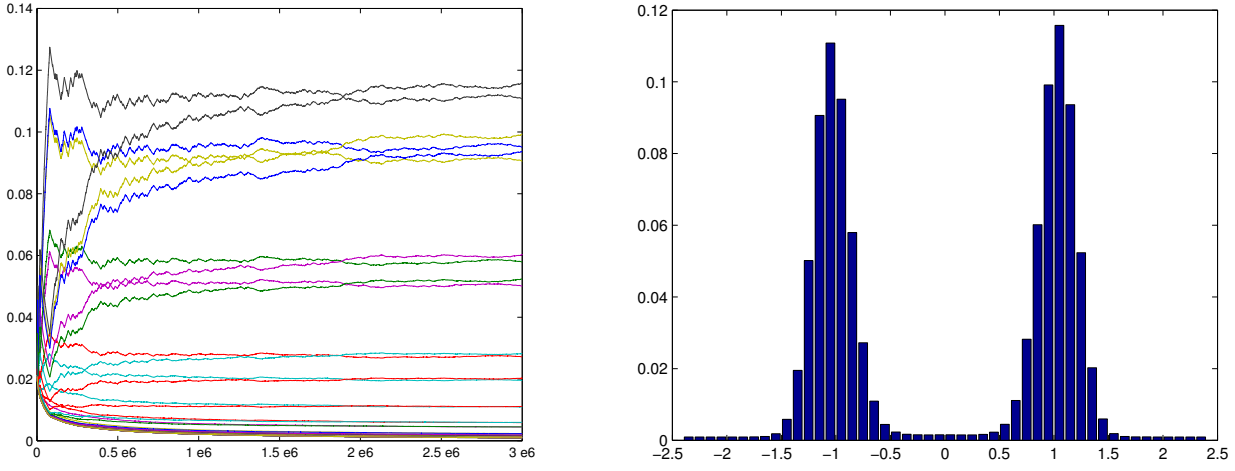


FIGURE: [left] The sequences $(\theta_k(i))_k$. [right] The limiting value $\theta_\star(i)$

Example 2 : Wang-Landau (3/4)

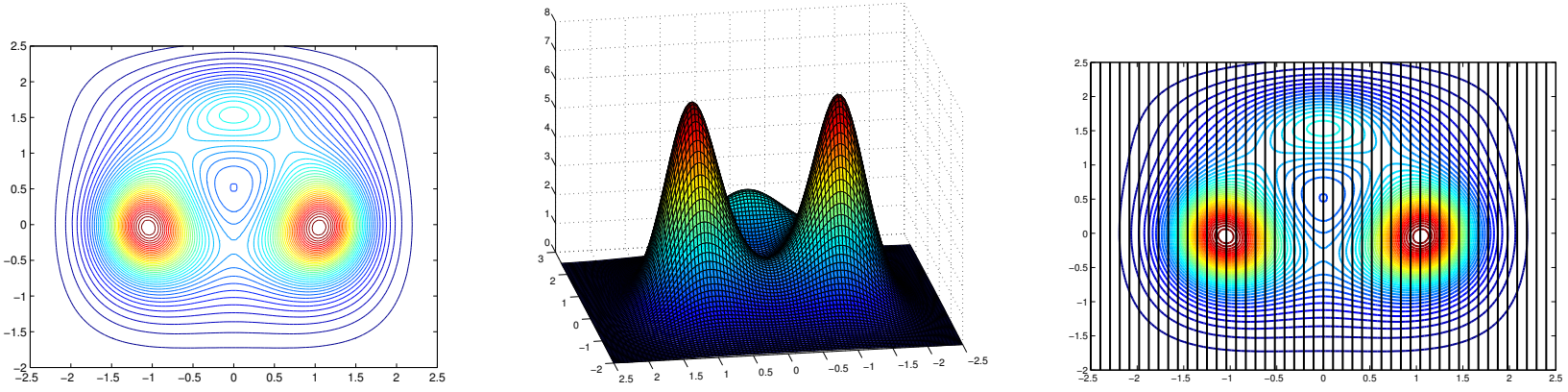


FIGURE: [left] level curves of π [center] Target density π [right] Partition of the state space

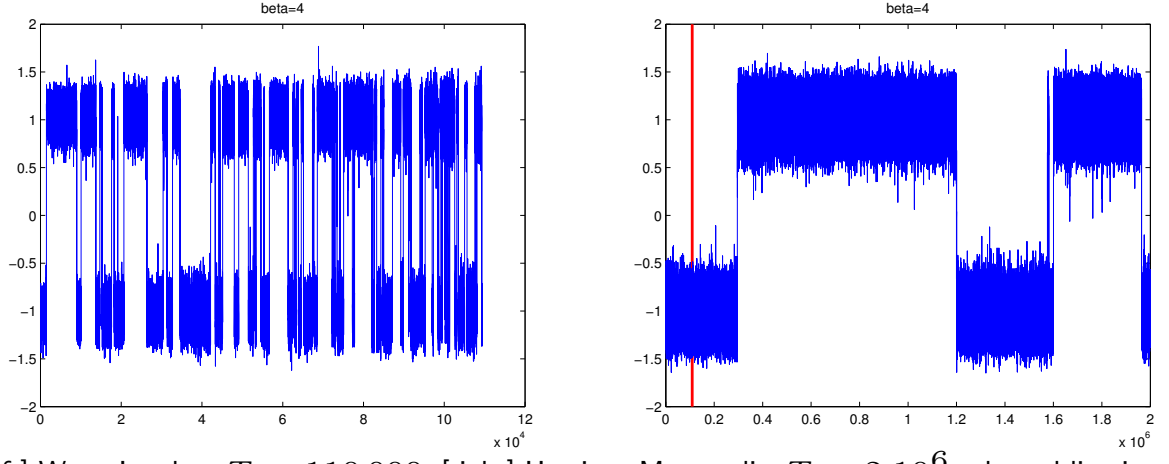


FIGURE: [left] Wang Landau, $T = 110\,000$. [right] Hastings Metropolis, $T = 2\ 10^6$; the red line is at $x = 110\,000$

Example 2 : Wang-Landau (4/4)

The general framework :

- Let π_θ be a distribution.
- Let P_θ be MCMC sampler with target distribution π_θ.
- For any θ : $\pi_\theta P_\theta = \pi_\theta$

The adaptive algorithm :

(i) Sample

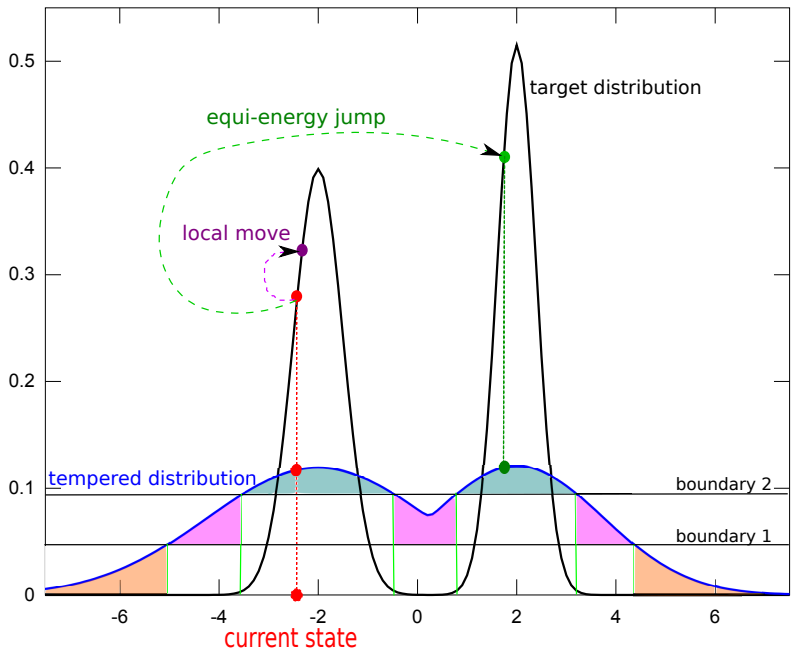
$$X_{k+1}|\mathcal{F}_k \sim P_{\theta_k}(X_k, \cdot)$$

(ii) Update the parameter θ_{k+1} by using θ_k, X_{k+1}.

Here, $\theta = (\theta(1), \cdots, \theta(d))$ is a probability on $\{1, \cdots, d\}$.

Example 3 : Equi-Energy (1/3)

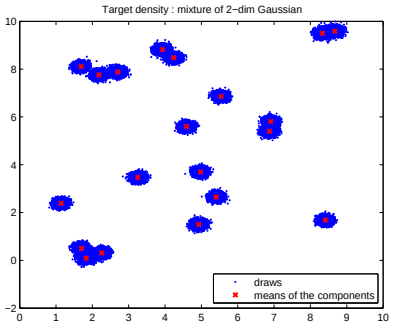
- Proposed by Kou et al. (2006) to sample multimodal target density π
- Based on an auxiliary process designed to admit $\pi^{1/T}$ $(T > 1)$ as target distribution.



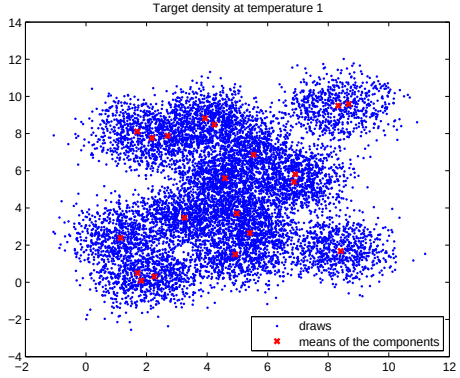
The transition kernel $X_k \to X_{k+1}$ is

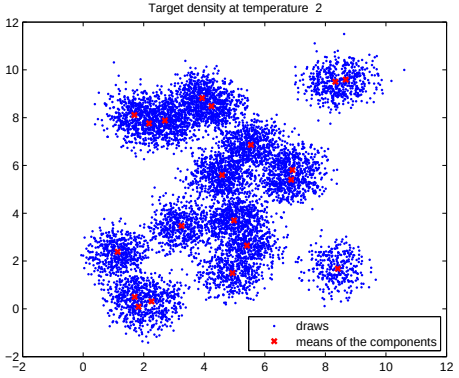
$$P_{\theta_k}(X_k, \cdot) = (1-\epsilon) \underbrace{Q(X_k, \cdot)}_{\text{MCMC with target } \pi} + \epsilon \underbrace{\tilde{Q}_{\theta_k}(X_k, \cdot)}_{\substack{\text{kernel depending on} \\ \text{the empirical distribution } \theta_k \text{ of the auxiliary process}}}$$

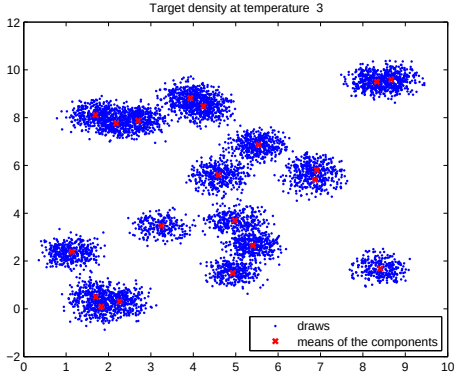
Example 3 : Equi-Energy (2/3)

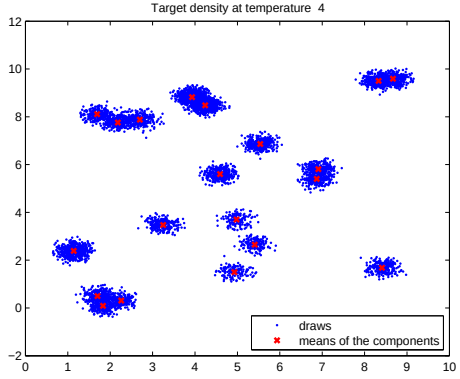


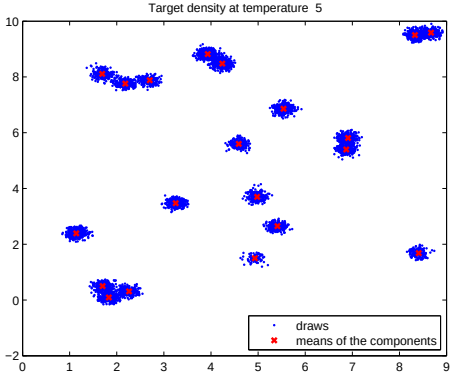
- target density : $\pi = \sum_{i=1}^{20} \mathcal{N}_2(\mu_i, \Sigma_i)$
- 5 processes with target distribution π^{1/T_k} $(T_K = 1)$

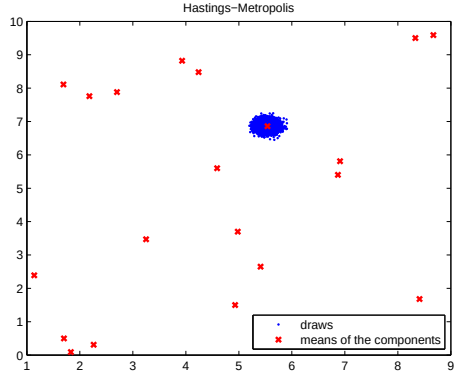












Example 3 : Equi-Energy (3/3)

The general framework :

- Let P_θ be the kernel associated to a EE-transition when the equi-energy jump uses a point sampled under the distribution θ.
- Under assumptions, for any θ : $\exists \pi_\theta$ s.t. $\pi_\theta P_\theta = \pi_\theta$.

The adaptive algorithm :

(i) Sample

$$X_{k+1} | \mathcal{F}_k \sim P_{\theta_k}(X_k, \cdot)$$

(ii) Update the distribution θ_{k+1} by using θ_k and $(\text{auxiliary process})_{k+1}$.

Here, θ_k is an empirical distribution on $\mathbb{X}$.

Unfortunately $\cdots$

- Unfortunately, adaption can destroy the convergence.
- Consider the following adapted Markov chain.
  - Let $\theta \in (0, 1)$. A Markov chain with transition matrix

$$P_\theta = \begin{pmatrix} 1-\theta & \theta \\ \theta & 1-\theta \end{pmatrix}$$

converges to the stationary distribution $\pi = (1/2; 1/2)$.

Unfortunately $\cdots$

- Unfortunately, adaption can destroy the convergence.
- Consider the following adapted Markov chain.
  - Let $\theta \in (0,1)$. A Markov chain with transition matrix

$$P_\theta = \begin{pmatrix} 1-\theta & \theta \\ \theta & 1-\theta \end{pmatrix}$$

converges to the stationary distribution $\pi = (1/2; 1/2)$.

  - Fix $t_0, t_1 \in (0,1)$. Define an **adapted chain** as follows :

$$X_{k+1}|\mathcal{F}_k \sim \begin{cases} P_{t_0}(X_k, \cdot) & \text{if } X_k = 0 \\ P_{t_1}(X_k, \cdot) & \text{if } X_k = 1 \end{cases}$$
$$\equiv P_{\theta_k}(X_k, \cdot) \qquad \text{with } \theta_k = t_{X_k}.$$

Unfortunately ...

- Unfortunately, adaption can destroy the convergence.
- Consider the following adapted Markov chain.
  - Let $\theta \in (0,1)$. A Markov chain with transition matrix

$$P_\theta = \begin{pmatrix} 1-\theta & \theta \\ \theta & 1-\theta \end{pmatrix}$$

converges to the stationary distribution $\pi = (1/2; 1/2)$.

  - Fix $t_0, t_1 \in (0,1)$. Define an **adapted chain** as follows :

$$X_{k+1}|\mathcal{F}_k \sim \begin{cases} P_{t_0}(X_k, \cdot) & \text{if } X_k = 0 \\ P_{t_1}(X_k, \cdot) & \text{if } X_k = 1 \end{cases}$$
$$\equiv P_{\theta_k}(X_k, \cdot) \qquad \text{with } \theta_k = t_{X_k}.$$

  - Then, $(X_k)_k$ is a Markov chain, with transition matrix

$$\begin{pmatrix} 1-t_0 & t_0 \\ t_1 & 1-t_1 \end{pmatrix}$$

but it converges to the distribution $\tilde{\pi} \propto (t_1, t_0) \neq \pi$.

Ergodicity (1/2)

Roberts and Rosenthal (2007); F., Moulines and Priouret (2012)

$$\begin{aligned}\mathbb{E}\left[f(X_t)\right] - \pi_{\theta_\star}(f) = &\mathbb{E}\left[f(X_t) - \mathbb{E}\left[f(X_t)|\mathcal{F}_{t-\ell}\right]\right] \\ &+ \mathbb{E}\left[\mathbb{E}\left[f(X_t)|\mathcal{F}_{t-\ell}\right] - P^\ell_{\theta_{t-\ell}} f(X_{t-\ell})\right] \\ &+ \mathbb{E}\left[P^\ell_{\theta_{t-\ell}} f(X_{t-\ell}) - \pi_{\theta_{t-\ell}}(f)\right] \\ &+ \mathbb{E}\left[\pi_{\theta_{t-\ell}}(f) - \pi_{\theta_\star}(f)\right]\end{aligned}$$

Convergence when

- the first term is null
- the second term is small when **adaption is diminishing**
- the third term is small when the transition kernels $(P_\theta, \theta \in \Theta)$ are ergodic (enough), at a rate which is uniform (enough) in θ (**containment condition**)

Ergodicity (2/2)

The last term : $\mathbb{E}\left[\pi_{\theta_{t-\ell}}(f) - \pi_{\theta_\star}(f)\right]$

1. Case 1 : $\pi_\theta = \pi$ for any θ.
   ex. Adaptive Metropolis
2. Case 2 : explicit expression of π_θ.
   ex. Wang-Landau F., Jourdain, Kuhn, Lelièvre & Stoltz (2012)
3. Case 3 : NO expression of π_θ BUT we have an expression of P_θ.
   $\hookrightarrow$ F., Moulines & Priouret (2012) check if π_θ inherits the smooth-in-θ conditions on the kernel P_θ.
   ex. Equi-Energy sampler F., Moulines & Priouret (2012) and Schreck, F. & Moulines (2013)

Central Limit Theorem (1/2)

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

- In the case of a (non adaptive) Markov chain i.e. $P_\theta = P$ then the variance in the CLT is given by

$$\sigma^2(f) = \int \pi(dx) \ (\Lambda f)^2(x) - \left(\int \pi(dx) \ \Lambda f(x) \right)^2$$

where Λf is the solution to the Poisson equation $f - \pi(f) = \Lambda f - P\Lambda f$.

- For adapted and interacting MCMC, it is true that

$$\sigma^2(f) = \int \pi_{\theta_\star}(dx) \ (\Lambda_{\theta_\star} f)^2(x) - \left(\int \pi_{\theta_\star}(dx) \ \Lambda_{\theta_\star} f(x) \right)^2 \qquad ?$$

$\hookrightarrow$ Not always : adaption/interaction may introduce an additional term.

Central Limit Theorem (2/2)

recall : $X_{k+1}|\mathcal{F}_k \sim P_{\theta_k}(X_k, \cdot) \qquad \pi_{\theta_k} P_{\theta_k} = \pi_{\theta_k}$

and

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

- General conditions are provided by F., Moulines, Priouret, Vandekerkhove (2012)
- For Adaptive Metropolis : Saksman, Vihola (2010), F., Moulines, Priouret, Vandekerkhove (2012)
  - $\pi_\theta = \pi$ for any θ.
  - Step 1 : show that $\lim_n \theta_n = \theta_\star$ w.p.1
  - Step2 : NO additional term

$$\sigma^2(f) = \int \pi(dx) \ \left(\Lambda_{\theta_\star} f\right)^2 (x) - \left(\int \pi(dx) \ \Lambda_{\theta_\star} f(x) \right)^2$$

Central Limit Theorem (2/2)

recall : $X_{k+1}|\mathcal{F}_k \sim P_{\theta_k}(X_k, \cdot) \qquad \pi_{\theta_k} P_{\theta_k} = \pi_{\theta_k}$

and

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

- General conditions are provided by F., Moulines, Priouret, Vandekerkhove (2012)
- For Wang-Landau : F. Jourdain, Kuhn, Kelièvre & Stoltz (2012)
  - $\pi_\theta P_\theta = \pi_\theta$.
  - Step 1 : show that $\lim_n \theta_n = \theta_\star$ w.p.1
  - Step 2 : NO additional term

$$\sigma^2(f) = \int \pi_{\theta_\star}(dx) \ \left(\Lambda_{\theta_\star} f\right)^2(x) - \left(\int \pi_{\theta_\star}(dx) \ \Lambda_{\theta_\star} f(x) \right)^2$$

since the Stochastic Approximation update of θ_n implies that rapidly enough

$$\| \pi_{\theta_n}(f) - \pi_{\theta_\star}(f) \| \leq C \| \theta_n - \theta_\star \| \to 0$$

.

Central Limit Theorem (2/2)

recall : $X_{k+1}|\mathcal{F}_k \sim P_{\theta_k}(X_k, \cdot)$ $\pi_{\theta_k} P_{\theta_k} = \pi_{\theta_k}$

and

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

- General conditions are provided by F., Moulines, Priouret, Vandekerkhove (2012)
- For Equi-Energy : F., Moulines, Priouret & Vandekerkhove (2012)
  - $\pi_\theta P_\theta = \pi_\theta$.
  - Step 1 : show that $\lim_n \theta_n = \theta_\star$ in some sense (convergence of measures)
  - Step 2 : additional term

$$\sigma^2(f) = \int \pi_{\theta_\star}(dx) \ (\Lambda_{\theta_\star} f)^2(x) - \left(\int \pi_{\theta_\star}(dx) \ \Lambda_{\theta_\star} f(x) \right)^2 + \gamma^2(f)$$

where $\gamma^2(f)$ collects the fluctuations of the auxiliary process

$$n^{-1/2} \sum_{j=1}^{\lfloor nt \rfloor} \left(f(Y_j) - \theta_\star(f) \right) \xrightarrow{d} \gamma^2(f) B_t \qquad (B_t) \text{std Brownian}$$

Conclusion

- There exist tools in the literature to prove the validity of adaptive and interacting MCMC.
  Results on the asymptotic behavior of the algorithms.

- What about explicit rate of convergence, explicit control of errors after a fixed number of iterations ? How to define a measure of efficiency ?

Adaptive MCMC algorithms (survey)

Andrieu, C. and Robert, C. (2001). Controlled markov chain monte carlo methods for optimal sampling. Tech. Rep. 125, Cahiers du Ceremade.

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. Statistics and Computing 18 :343-373.

Atchade, Y. and Fort, G. and Moulines, E. and Priouret, P. (2011) Adaptive Markov chain Monte Carlo : Theory and Methods. Bayesian Time Series Models, Cambridge Univ. Press, Chapter 2, 33-53.

Atchade, Y.F. and Rosenthal, J.S. (2005). On adaptive Markov chain Monte Carlo algorithm. Bernoulli 11 :815-828.

Roberts, G. and Rosenthal, J. (2009). Examples of adaptive MCMC. J. Comp. Graph. Stat. 18 :349-367.

Rosenthal, J. S. (2009). MCMC Handbook, chap. Optimal Proposal Distributions and Adaptive MCMC. Chapman & Hall/CRC Press.

Convergence of adaptive and interacting MCMC
Roberts, G.O. and Rosenthal, J.S. Coupling and ergodicity of adaptive MCMC. J. Appl. Probab. 44 :458-475 (2007).

Fort, G., Moulines, E. and Priouret, P. Convergence of interacting MCMC : ergodicity and law of large numbers. Ann. Statist. 39 :3262-3289 (2012)

Fort, G., Moulines, E., Priouret, P. and Vandekerkhove, P. Convergence of interacting MCMC : Central Limit Theorem. Bernoulli (2013).

Latuszynski, K. and Rosenthal, J.S. The containment condition and AdapFail algorithms. Submitted (2013).

Convergence of stochastic approximation scheme
A. Benveniste, M. Metivier and P. Priouret. Adaptive algorithms for Stochastic Approximations. Springer-Verlag (1987).

C. Andrieu, E. Moulines and P. Priouret. Stability of stochastic approximation under verifiable conditions. SIAM Journal on Control and Optimisation 44 :283-312 (2005).

G. Fort. Central Limit Theorems for Stochastic Approximation with controlled Markov chain dynamics. Submittted (2013).

Convergence of Adaptive Metropolis
Saksman, H. and Vihola, M. On the ergodicity of the adaptive Maetropolis algorithm on unbounded domains. Ann. Appl. Probab. (2010) 20 2178-2203.

Fort, G., Moulines, E. and Priouret, P. Convergence of interacting MCMC : ergodicity and law of large numbers. Ann. Statist. 39 :3262-3289 (2012)

Fort, G., Moulines, E., Priouret, P. and Vandekerkhove, P. Convergence of interacting MCMC : Central Limit Theorem. Bernoulli (2013).

Convergence of the Equi-Energy sampler
Hua, X. and Kou, S.C. Convergence of the Equi-Energy Sampler and Its Application to the Ising Model (2011) Stat. Sin. 24 :1687-1711.

Fort, G., Moulines, E. and Priouret, P. Convergence of interacting MCMC : ergodicity and law of large numbers. Ann. Statist. 39 :3262-3289 (2012)

Fort, G., Moulines, E., Priouret, P. and Vandekerkhove, P. Convergence of interacting MCMC : Central Limit Theorem. Bernoulli (2013).

A. Schreck, G. Fort and E. Moulines. Adaptive Equi-energy sampler : convergence and illustration. Accepted in Transactions on Modeling and Computer Simulation (2012).

Methodology and Convergence analysis of Wang-Landau

F. Liang. A general Wang-Landau algorithm for Monte Carlo computation. J. Am. Stat. Assoc. 100 :1311-1327 (2005).

F. Liang, C. Liu and R.J. Carroll. Stochastic approximation in Monte Carlo computation. J. Am. Stat. Assoc. 102 :305-320 (2007).

Y. Atchadé and J.S. Liu. The Wang-Landau algorithm for Monte Carlo computation in general state space. Stat. Sinica, 20(1) :209-233 (2010).

Application of Wang-Landau to Statistics. Convergence results (on the samples $(X_t)_t$) under the assumption that the algorithm is "stable"

G. Fort, B. Jourdain, E. Kuhn, T. Lelièvre and G. Stoltz. Convergence of the Wang Landau algorithm. In revision (2013).

Sufficient conditions for (i) the stability, the convergence and the rate of convergence of the sequence of weights $\{\theta_n\,,\, n \geq 0\}$; (ii) the convergence of the sampling method.

G. Fort, B. Jourdain, E. Kuhn, T. Lelièvre and G. Stoltz. Efficiency of the Wang Landau algorithm. In revision (2013).

Discussion on the efficiency of the Wang-Landau algorithm

L. Bornn, P. Jacob, P. Del Moral and A. Doucet. An Adaptive Wang-Landau Algorithm for Automatic Density Exploration. To appear in Journal of Computational and Graphical Statistics (2013).

New methods for (i) adaptive binning strategy to automate the difficult task of partitioning the state space, (ii) the use ofinteracting parallel chains to improve the convergence speed and use of computational resources, and (iii) the use of adaptive proposal distributions.

P. Jacob and R. Ryder. The Wang-Landau algorithm reaches the flat histogram criterion in finite time. To appear in Ann. Appl. Probab. (2013).

The linearized version of the update of the weight vector θ_t satisfies in finite time the uniformity criterion required in the original Wang-Landau algorithm. This is not guaranteed for some non-linear update.