Adaptive and Interacting MCMC algorithms

Eric MOULINES

Telecom Paris Tech
CNRS - LTCI

Joint work with **G. FORT** & S. LE CORFF (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 lack of convergence may be difficult to diagnose.
3. In the last 10 years, several algorithms have been proposed to **increase** the sampling efficiency of the MCMC, without requiring much additional user supervision.
4. The common idea is to improve the sampling strategy by **learning** from the past simulations.

Outline of the talk

1. **Algorithm design**
   - Adaptive Markov chain : a single chain whose kernel is gradually modified
   - Interacting Markov chains : multiple chains which interact
2. **Some numerical examples**
3. **Convergence of the algorithms**

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 $\bar{q}$ is **symmetric** (i.e. $\bar{q}(z) = \bar{q}(-z)$)
- 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

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

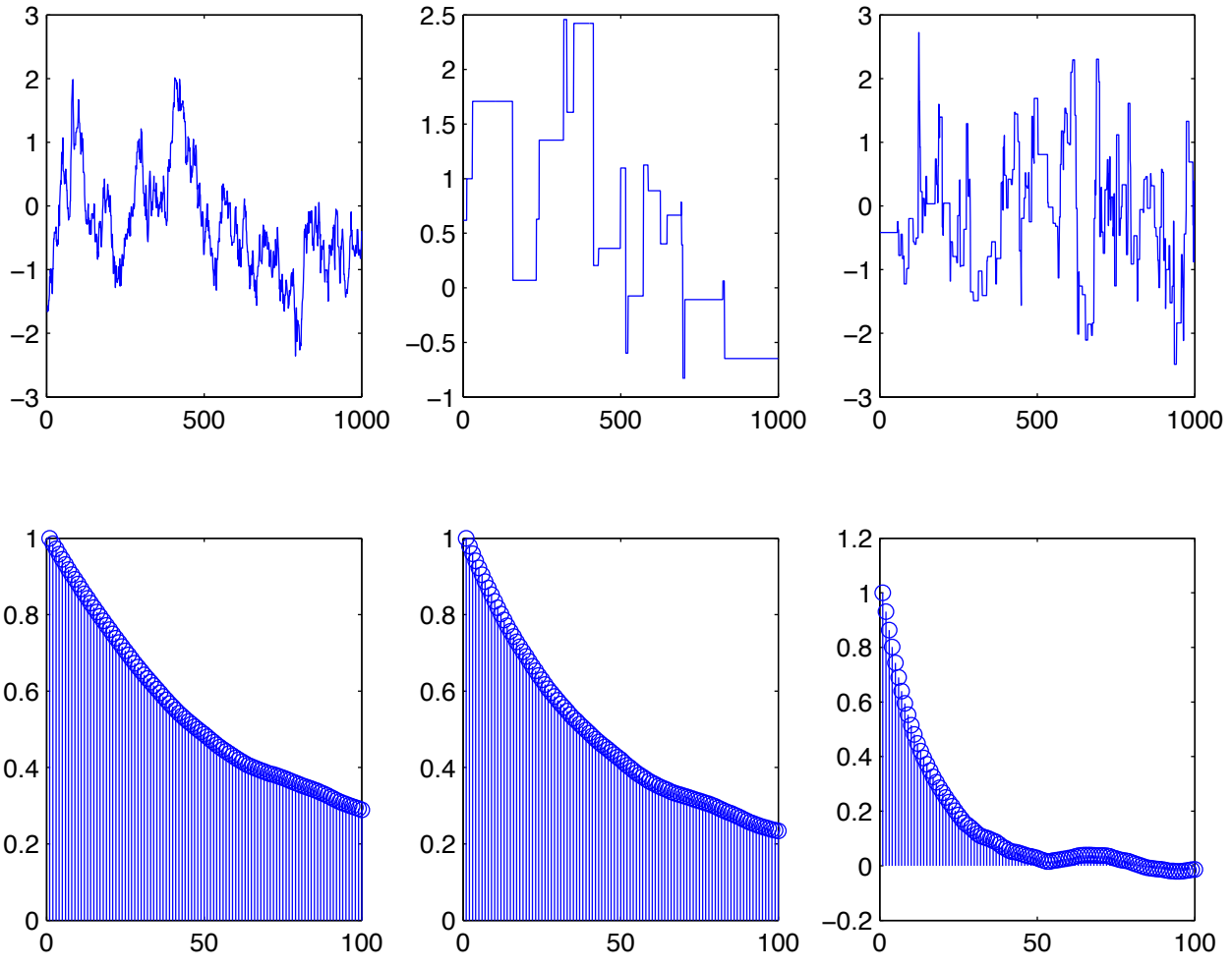
- ... biased random walk where some moves get rejected.

Influence of the scaling

- If the variance is either **too small** or **too large**, then the convergence rate of the Markov chain is slow :
  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

Optimal Scaling of the RWM

- A useful idea to get a better understanding of the influence of scaling is to consider a **high-dimensional** limit, *i.e.* the state space $\mathsf{X} = \mathbb{R}^d$ where we let the dimension $d \to \infty$.
- Under appropriate assumptions, each coordinate of the Markov chain $\{X_{k,i}^{(d)}\}_{i=1}^{d}$ converges to a diffusion limit.
- The choice of an optimal scaling then translates into the optimization of the limiting diffusion speed, which is rather easy to handle.

Diffusive Limits

- **Stationary distribution :** $\pi^{(d)}(x_1, \ldots, x_d) = \prod_{i=1}^{d} f(x_i)$ on $\mathbb{R}^d$ $(d \to \infty)$
- **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 becomes independent from the other components which globally act as a random environment.

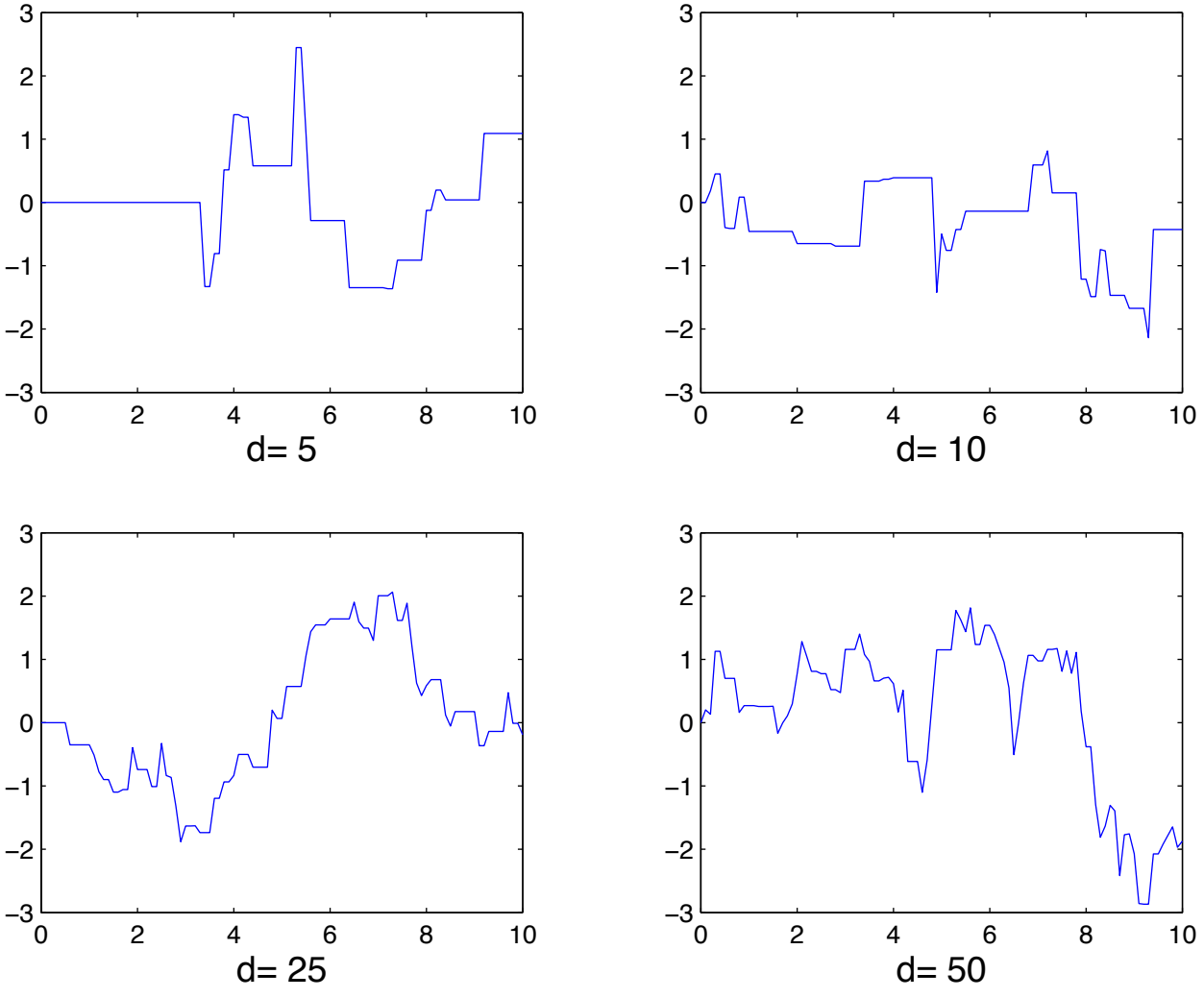


FIGURE: Diffusive limits for different values of d

Diffusive Limits

$Z^{(d)} \Rightarrow_d Z$ in the Skorokhod space, where Z is a solution the Langevin SDE

$$\mathrm{d}Z_t = v^{1/2}(\theta)\mathrm{d}B_t + (1/2)v(\theta)\nabla \log f(Z_t)\mathrm{d}t$$
$$v(\theta) = \theta^2 \tau^{(\infty)}\left[\theta, I(f)\right]$$

where,

$$\tau^{(\infty)}\left[\theta, I(f)\right] = \lim_{d\to\infty} \tau^{(d)}(\theta)$$

is the limit of the acceptance rate in stationarity,

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

with $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ and

$$I(f) = \int \left[\left(\frac{\mathrm{d}\log f(x)}{\mathrm{d}x}\right)'\right]^2 \mathrm{d}x\ .$$

Diffusion speed

- $v(\theta) = 2\theta^2 \tau^{(\infty)}\left[\theta, I(f)\right]$ is the **speed** of the limiting diffusion : $Z_t = \tilde{Z}_{v(\theta)t}$ where $\{\tilde{Z}_t\}$ is a solution of the Langevin SDE

$$\mathrm{d}\tilde{Z}_t = \mathrm{d}B_t + (1/2)\nabla \log f(\tilde{Z}_t)\mathrm{d}t \; .$$

- Optimizing the scale amounts to find θ which maximizes the diffusion speed.

Diffusion speed optimization

- The limiting acceptance rate is given
$$\tau^{(\infty)}[\theta] = 2\Phi\left(\theta\frac{\sqrt{I(f)}}{2}\right) \Longleftrightarrow \theta = \frac{2}{\sqrt{I(f)}}\Phi^{-1}\left(\tau^{(\infty)}[\theta]/2\right) .$$
- Since $v(\theta) = \theta^2\tau^{(\infty)}[\theta]$ the speed may be rewritten as a function of the mean acceptance rate in stationarity
$$v(\theta) \propto w\left[\tau^{(\infty)}(\theta)\right] \quad w : \tau \mapsto \tau\Phi^{-1}(\tau/2) .$$
- The speed is maximized if the scale is chosen so that $\tau^{(\infty)}[\theta_\star]$, where $\bar{\tau}$ is the maximum of w.
- The optimum value of the acceptance rate may be shown to be $\bar{\tau} \approx 0.234$...

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 scale θ therefore amounts to solve

$$h(\theta) \stackrel{\text{def}}{=} \iint \left\{1 \wedge \frac{\pi(y)}{\pi(x)}\right\} \frac{1}{\theta} q\left(\frac{y-x}{\theta}\right) \pi(x)\mathrm{d}x\mathrm{d}y - \bar{\tau} = 0,$$

- Under appropriate 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 !
- **Suggest to use a stochastic approximation procedure to adapt the scale θ.**

Adaptive Scaling Metropolis Algorithm

- Proposition & Accept/Reject

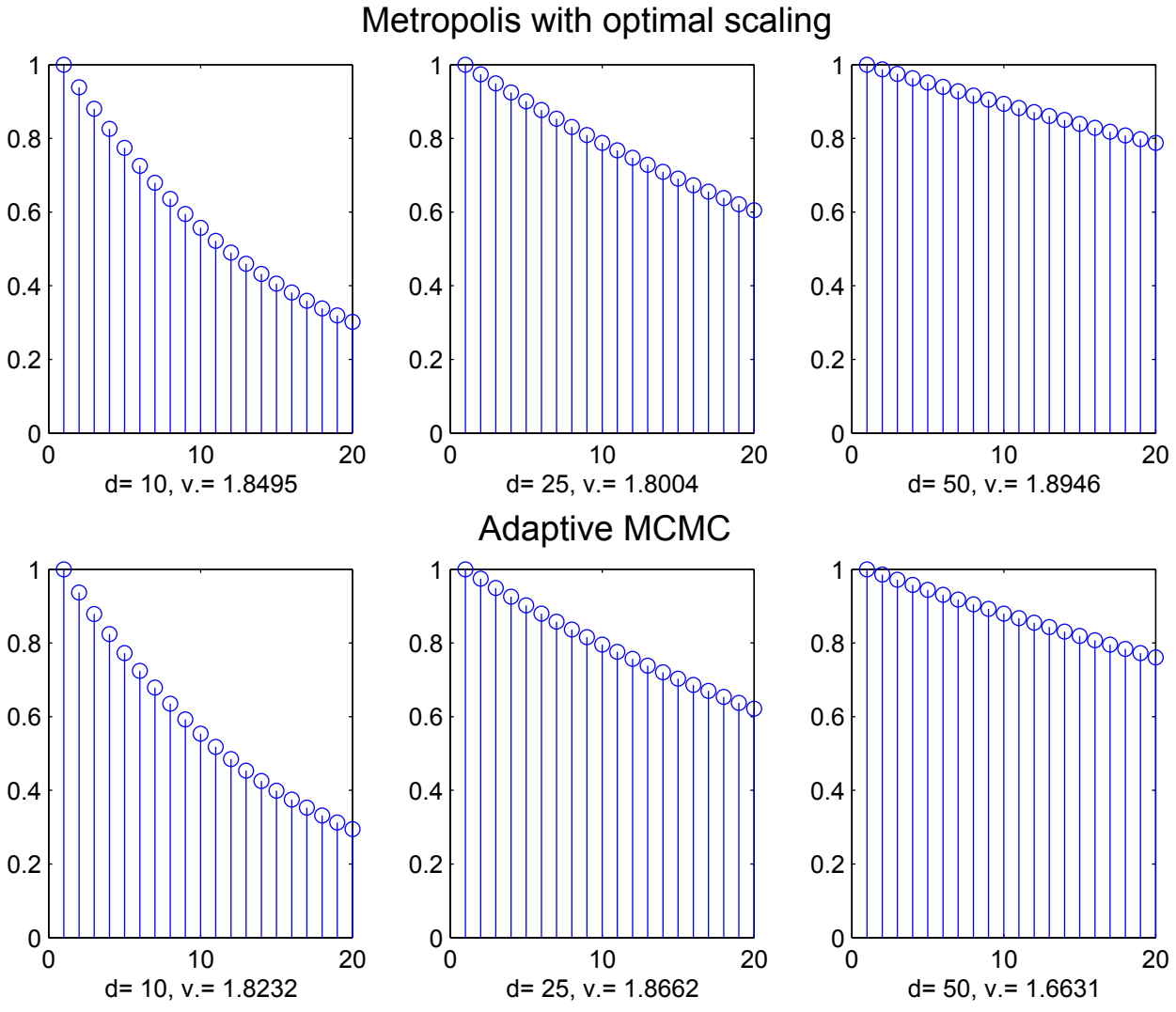
$$Y_{k+1} = X_k + \theta_k \mathcal{N}(0, \mathrm{Id})$$

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

- 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 with optimal scaling
d= 10, v.= 1.8495
d= 25, v.= 1.8004
d= 50, v.= 1.8946
Adaptive MCMC
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 the covariance but the maximal achievable speed is strongly affected.
- **Idea** : adapt the scale and the covariance of the proposal.

Adaptive MCMC with multidimensional scaling

1. Simulate

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

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

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

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

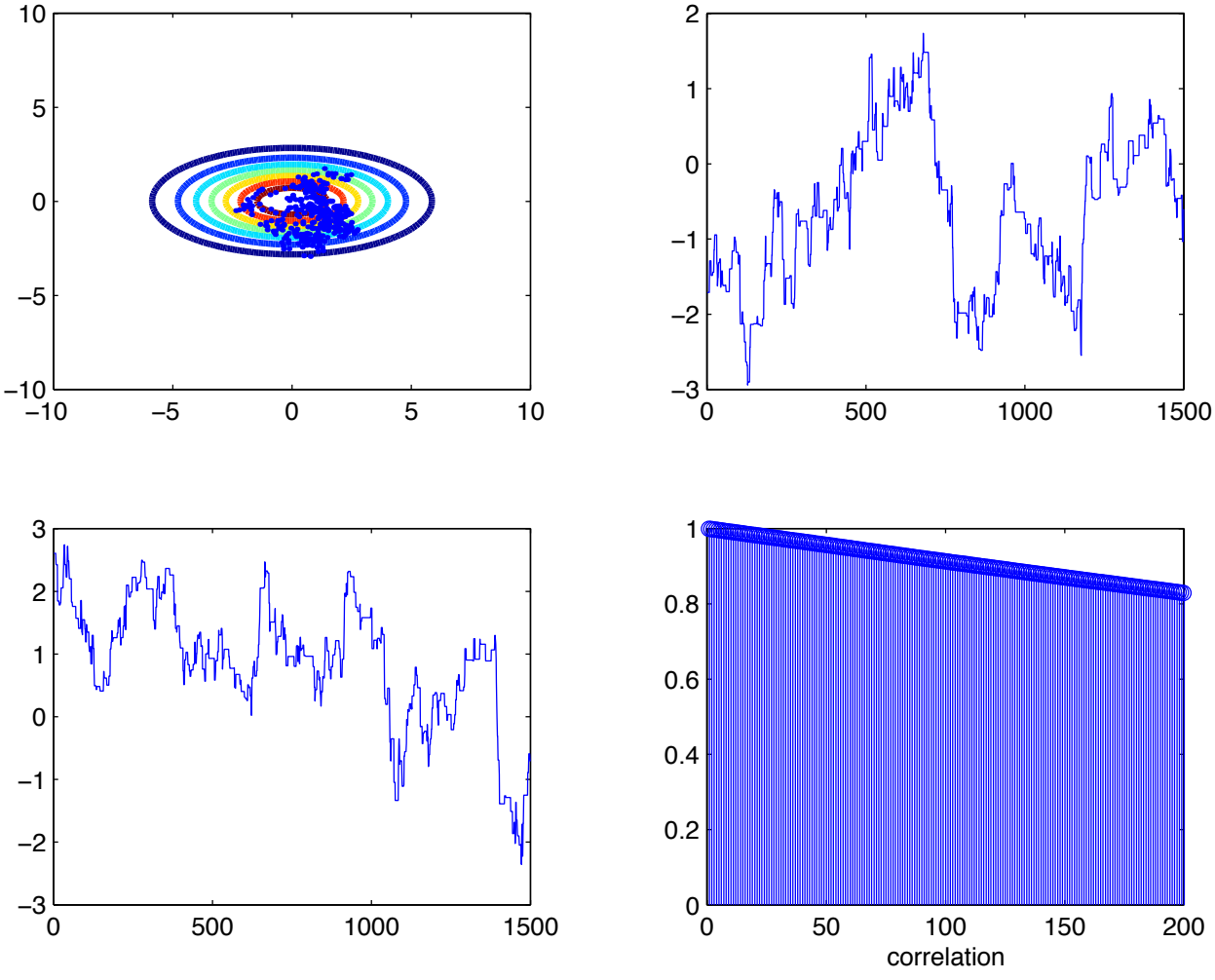


FIGURE: $d = 12$, $\pi \sim \mathcal{N}(0, \Gamma)$, $\mathrm{cond}(\Gamma) \approx 100$, $q \sim \mathcal{N}(0, (2.32^2/d)\,\mathrm{I})$

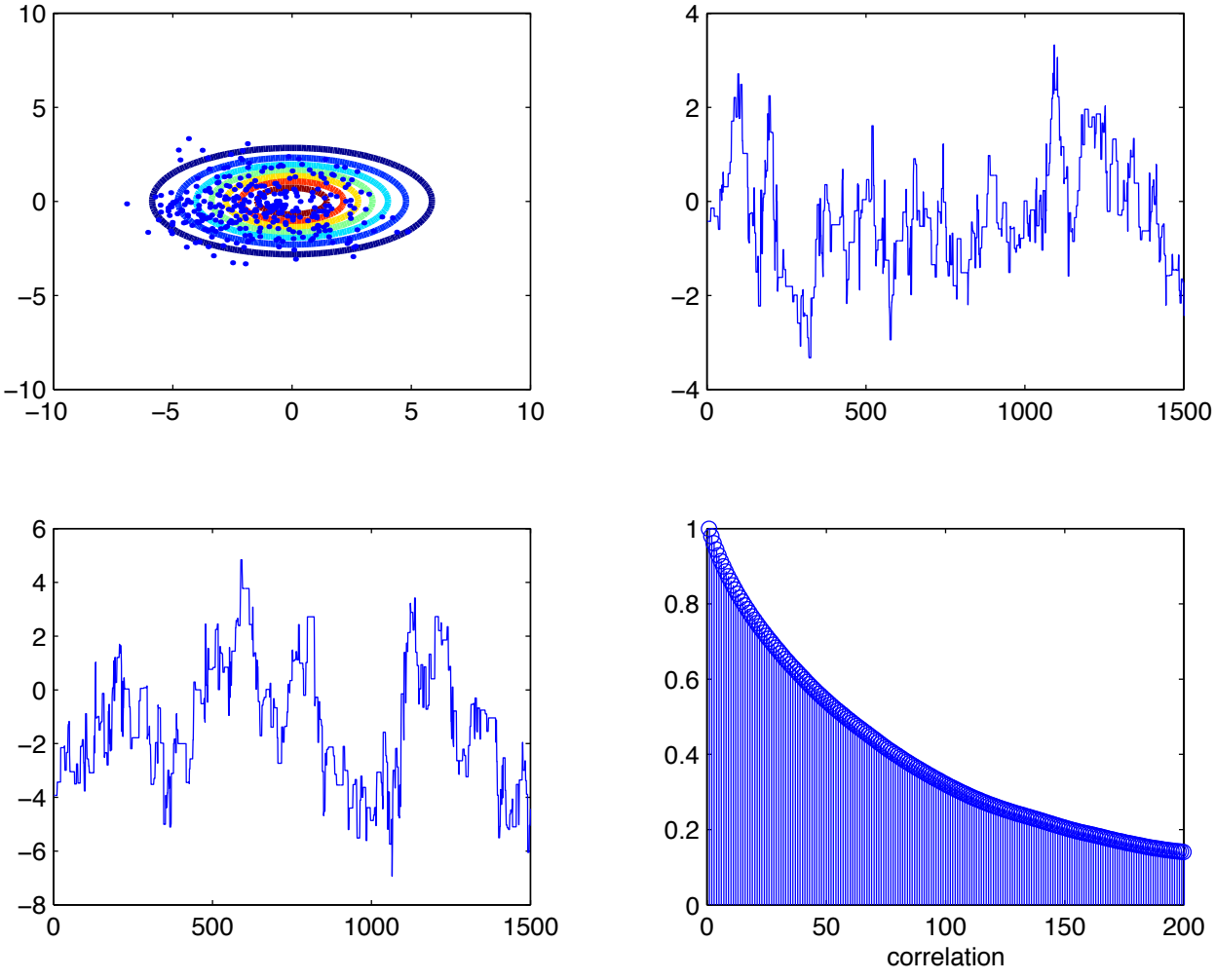


FIGURE: $d = 12$, $\pi \sim \mathcal{N}(0, \Gamma)$, $\mathrm{cond}(\Gamma) \approx 100$, $q \sim \mathcal{N}(0, (2.32^2/d)\,\Gamma)$

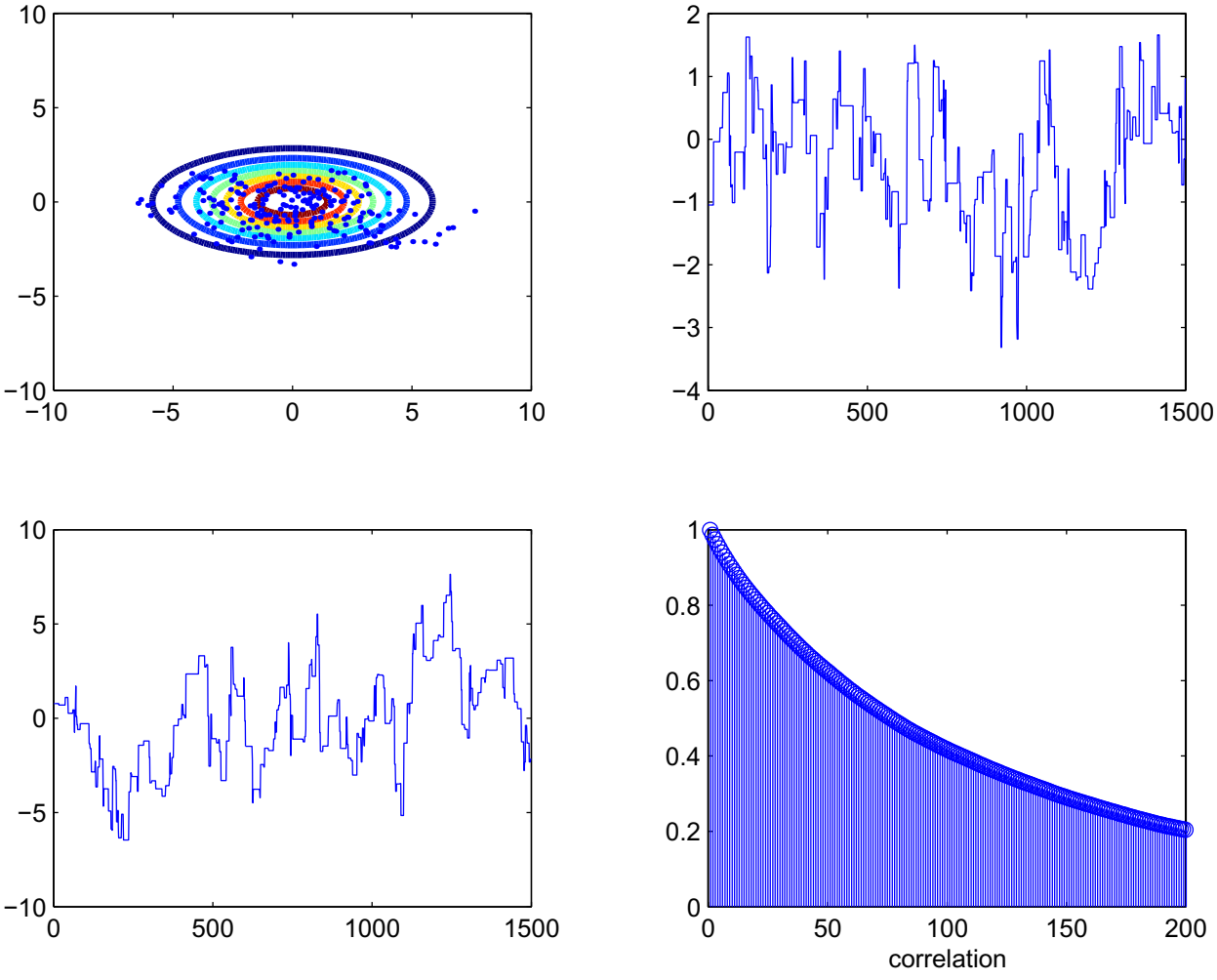


FIGURE: $d = 12$, $\pi \sim \mathcal{N}(0, \Gamma)$, $\mathrm{cond}(\Gamma) \approx 100$, $q \sim \mathcal{N}(0, \sigma_k \Gamma_k)$, with adaptive multidimensional scaling

Tempering

- The adaptive Metropolis-Hastings algorithm outlined above can run into difficulties if the target probability distribution is multi-modal.
- The MCMC can become trapped in a local mode and fail to fully explore other modes which are significant.
- This problem is very similar to the one encountered in finding a **global minimum** in nonlinear optimisation. One solution to that problem was to use **simulated annealing** by introducing a **temperature parameter**.
- The analogous process applied to drawing samples from a target probability distribution is often referred to as **tempering** : instead of **cooling down** to make the distribution **sharper and sharper**, we rather **heating up** the distribution to make it **flatter and flatter**...

Parallel tempering

- In **parallel tempering** algorithm by Geyer (1991) is to run parallel Metropolis sampling at different **temperatures** $T_1 \geq T_2 \geq \cdots \geq T_K = 1$, with target distributions $\left\{\pi^{1/T_k}\right\}_{k=1}^{K}$.
- At intervals, a pair of adjacent level is chosen and a proposal made to swap their states. If the swap is accepted then these states are interchanged.
- The acceptance probability for the swap between the state at temperature T_{k-1} and T_k ($k \in \{2, \ldots, K\}$) is computed to ensure that the joint states of all the parallel chains is reversible with respect to the tensor product $\pi^{1/T_1} \otimes \cdots \otimes \pi^{1/T_K}$ of the heated up probability :

$$\alpha_k\left(x^{(k-1)}, x^{(k)}\right) = 1 \wedge \frac{\pi^{1/T_{k-1}}\left(x^{(k)}\right)\pi^{1/T_k}\left(x^{(k-1)}\right)}{\pi^{1/T_{k-1}}\left(x^{(k-1)}\right)\pi^{1/T_k}\left(x^{(k)}\right)} .$$

Parallel tempering

- This swap allows for an exchange of information across the population of parallel simulations.
- In the higher temperature simulations, radically different configurations can arise.
- By making exchanges, we can capture and improve configurations by putting them into lower temperature simulations.
- **Drawback** : The temperature levels should be close enough to achieve a significant acceptance probability for a swap.

Interacting Tempering

- The Interacting Tempering Algorithm exploits the parallel tempering idea : the algorithm runs several chains at different temperatures.
- The idea is to replace an **instantaneous swap** by an **interaction** with the whole past of a neighboring process.
- **Idea** : At time n, find in the past samples of the chain $X_\star^{(k-1)} \in \{X_0^{(k-1)}, \ldots, X_n^{(k-1)}\}$ run at temperature T_{k-1} a state such that the probability of accepting the move

$$\frac{\pi^{1/T_{k-1}}\left(X_n^{(k)}\right)\pi^{1/T_k}\left(X_\star^{(k-1)}\right)}{\pi^{1/T_{k-1}}\left(X_\star^{(k-1)}\right)\pi^{1/T_k}\left(X_n^{(k)}\right)} .$$

is large enough.

Interacting Tempering (at temperature T_i)

- a transition kernel $P^{(k)}$ with stationary distribution π^{1/T_k} : $\pi^{1/T_k} P^{(k)} = \pi^{1/T_k}$ (typically, a MH algorithm run with the target distribution π^{1/T_k}).
- a probability of interaction $\epsilon \in (0,1)$

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

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

Interacting Tempering

with probability ϵ,

- **select** a state in $X_\star^{(k-1)} \in \left\{X_\ell^{(k-1)}\right\}_{\ell=0}^n$ with probability $\left\{g(X_n^{(k)}, X_\ell^{(k-1)})\right\}_{\ell=0}^n$;
- **accept** the proposal with probability $\alpha_k(X_n^{(k)}, X_\star^{(k-1)})$

$$P_{\theta_n^{(k-1)}}(X_n^{(k)}, A) = (1-\epsilon)P^{(k)}(X_n^{(k)}, A) + \epsilon \left\{ \int_A \theta_n^{(k-1)}(\mathrm{d}y)\alpha_k(X_n^{(k)}, y) + \mathbb{1}_A(X_n^{(k)}) \int \theta_n^{(k-1)}(\mathrm{d}y) \left\{1 - \alpha_k(X_n^{(k)}, y)\right\} \right\}$$

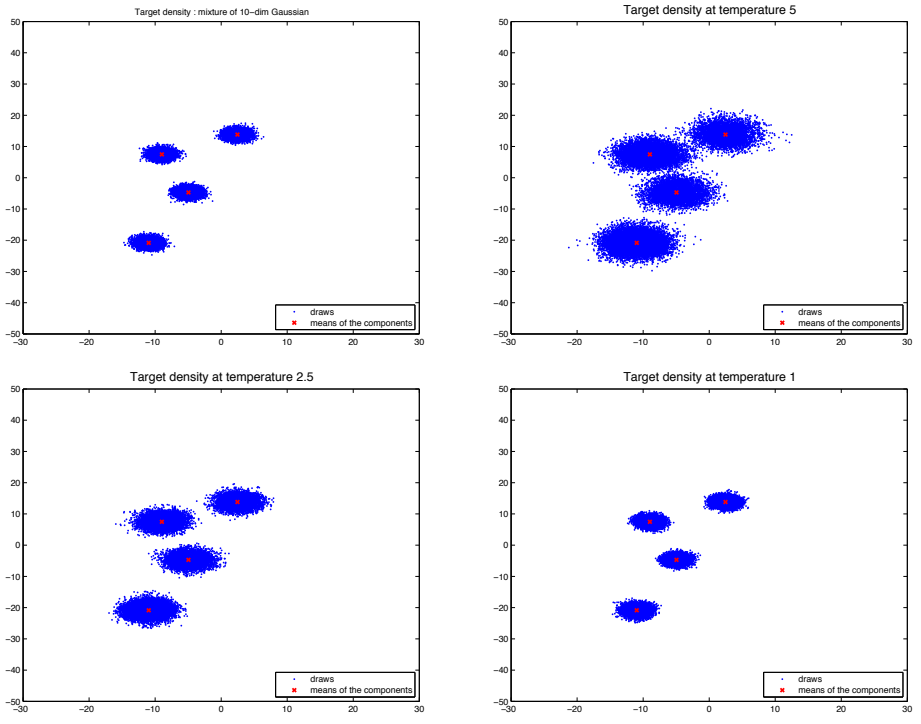
where $\theta_n^{(k-1)}(\mathrm{d}y) = \frac{1}{n+1}\sum_{\ell=1}^n \delta_{X_\ell^{(k-1)}}(\mathrm{d}y)$ and

$$\alpha_k(x,y) = \frac{g(x,y)}{\int \theta_n^{(k-1)}(\mathrm{d}y) g(x,y)} \left(1 \wedge \frac{\pi^{1/T_k}(y)\pi^{1/T_{k-1}}(x)}{\pi^{1/T_{k-1}}(y)\ \pi^{1/T_k}(x)}\right) .$$

An example

1. Mixture of Gaussians : 4 components in dimension 10.
2. Dimension : $d = 10$ (only the first two components are shown)
   Interactions : 5 %
3. Temperatures : 50,40,30,25,20,15,10,5,2.5,1
4. 50 Energy rings (adapted from the empirical quantiles)
5. Basic Kernel : random walk Metropolis with covariance $(4/d) * \mathrm{I}$
   (optimally adapted to individual components).

Interactions : 5 %



A General Framework

- Let $(\Theta, \mathcal{T})$ be a measurable space and $(\mathsf{X}, \mathcal{X})$ a general state space.
- Let $(P_\theta, \theta \in \Theta)$ be a collection of Markov transition kernels indexed by $\theta \in \Theta$, which can be either **finite** or **infinite dimensional** (e.g. an empirical distribution).
- For each $\theta \in \Theta$, P_θ admits a single probability distribution π_θ : $\pi_\theta = \pi_\theta P_\theta$.
- Consider a $\mathsf{X} \times \Theta$-valued process $\{(X_n, \theta_n), n \geq 0\}$ on a filtered probability space $(\Omega, \mathcal{A}, \{\mathcal{F}_n, n \geq 0\}, \mathbb{P})$ such that $\{(X_n, \theta_n), n \geq 0\}$ is $\mathcal{F}_n$-adapted and for any bounded measurable function f

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

Problems

- **Problem :** Find conditions such that :
  1. **Ergodicity :** $\lim_{n\to\infty} \mathbb{E}\left[f(X_n)\right] = \pi(f)$ where π is the **target distribution**.
  2. **Strong Law of Large Numbers** : $\lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} f(X_k) \to \pi(f)$ $\mathbb{P}$-a.s.
- **Problem :** $\{X_k\}$ is **not** a Markov Chain.
- **Existing results :**
  1. Adaptive Markov Chains : Andrieu, Moulines (2006), Roberts and Rosenthal (2007), Atchadé and Fort (2009)
  2. Interacting Markov Chains : Del Moral and Miclo (2004), Andrieu, Del Moral, Doucet, Jasra (2006,2007,2010), Del Moral and Doucet (2009), Bercu, Del Moral and Doucet (2009)

Error decomposition

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

Error decomposition

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

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

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

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

Error decomposition

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

$\hookrightarrow$ [B] condition on the adaptation mechanism

Error decomposition

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

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

Error decomposition

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

The conditions can be weakened by replacing N by $r(n)$. This allows to consider situations where the ergodicity constant of the Markov kernels get worse when θ approaches the parameter set boundary

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

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

Result

[Fort et al. 2010]

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

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

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

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

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

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

Law of large numbers for adaptive MCMC samplers

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

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

Sketch of the proof

We write

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

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

Sketch of the proof

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

For the first term, Tool : Poisson equation so that

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

Sketch of the proof

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

For the first term, Tool : Poisson equation so that

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

- Martingale increments : $\hookrightarrow$ [B] moment conditions : for some $\alpha > 1$,

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

Sketch of the proof

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

For the first term, Tool : Poisson equation so that

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

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

Result

[Fort et al. 2010]

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

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

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

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

C. (Strengthened diminishing adaptation)

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

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

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

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

Result

[Fort et al. 2010]

A. (Ergodicity of the transition kernels)

B. X is Polish

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

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

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

Application to the convergence of adaptive and interacting MCMC algorithms

Ergodicity criteria : checked in practice by

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

Application to the convergence of adaptive and interacting MCMC algorithms

Ergodicity criteria : checked in practice by

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

Diminishing adaptation : checked in practice by

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

Application to the convergence of adaptive and interacting MCMC algorithms

Ergodicity criteria : checked in practice by

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

Diminishing adaptation : checked in practice by

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

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

Adaptive MCMC

We prove

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

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

Convergence of the (simplified) Equi-Energy sampler

We prove

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

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

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