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 of the talk

1. **Algorithm design**
   - Adaptive Markov chain: a single chain whose kernel is gradually modified
   - Interacting Markov chains: multiple chains with interactions
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 $Z_{k+1} \sim_{\text{i.i.d.}} \bar{q}$, where $\bar{q}$ is **symmetric** (i.e. $\bar{q}(z) = \bar{q}(-z)$)

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

- **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

Optimal Scaling of the RWM

- A useful idea to understand MCMC is to consider some sort of limits.
- **high-dimensional** limit, *i.e.* the state space $\mathsf{X} = \mathbb{R}^T$ where we let the dimension $T \to \infty$, are in general useful.
- Under appropriate assumptions, each coordinate of the Markov chain $\{X_{k,i}^{(T)}\}_{i=1}^{T}$ converges to a diffusion limit.
- The choice of an appropriate scale then translates into the optimization of the limiting diffusion speed.

Diffusive Limits

- **Stationary distribution:** $\pi^{(T)}(x_1, \ldots, x_T) = \prod_{i=1}^{T} f(x_i)$ on $\mathbb{R}^T$ $(T \to \infty)$
- **Metropolis proposal**: $q_\theta^{(T)}(x_1, \ldots, x_T) \sim \mathcal{N}\left(0, (\theta^2/T)\mathrm{I}_T\right)$... with variance decreasing as $1/T$.
- **Interpolated process**: $Z_t^{(T)} = X_{[tT],1}^{(T)}$... we consider a single component and we speed up the time scale by T.

Diffusive Limits

$Z^{(T)} \Rightarrow_d Z$ where Z solves the Langevin SDE with limiting speed $v(\theta)$

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

where,

$$\tau^{(\infty)}(\theta) = \lim_{T\to\infty} \tau^{(T)}(\theta)$$

is the limit of the acceptance rate in stationarity.

Diffusion speed optimization

- The limiting speed $v(\theta) = \theta^2 \tau^{(\infty)}(\theta)$ may be rewritten as a function of the mean acceptance rate in stationarity

$$v(\theta) \propto w\left[\tau^{(\infty)}(\theta)\right] \qquad w : \tau \mapsto \tau \Phi^{-1}(\tau/2) \ .$$

- The speed is maximized if the scale is chosen so that $\tau^{(\infty)}\left[\theta_\star\right]$, 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.
- One major disadvantage of the diffusion limit work is its reliance on asymptotics in the dimensionality of the problem... and the target distribution should be simple enough to obtain interpretable limiting results.
- Extensions and generalisations of this result can be found in (Roberts and Rosenthal, 2001) and (Bedard, 2007), (Pillai, Stuart, 2009), introducing some forms of dependence between the components
- Other algorithms can benefit from this analysis: **multiple try Metropolis** and **Delayed Rejection** algorithms have been analysed in (Bedard, Douc, Moulines, 2010a & 2010b)... conclusions are non trivial and suggest way to design adaptive algorithms.

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 (see Andrieu, Robert, (2001), Vihola (2010)).

Adaptive Scaling Metropolis Algorithm

- Proposition & Accept/Reject

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

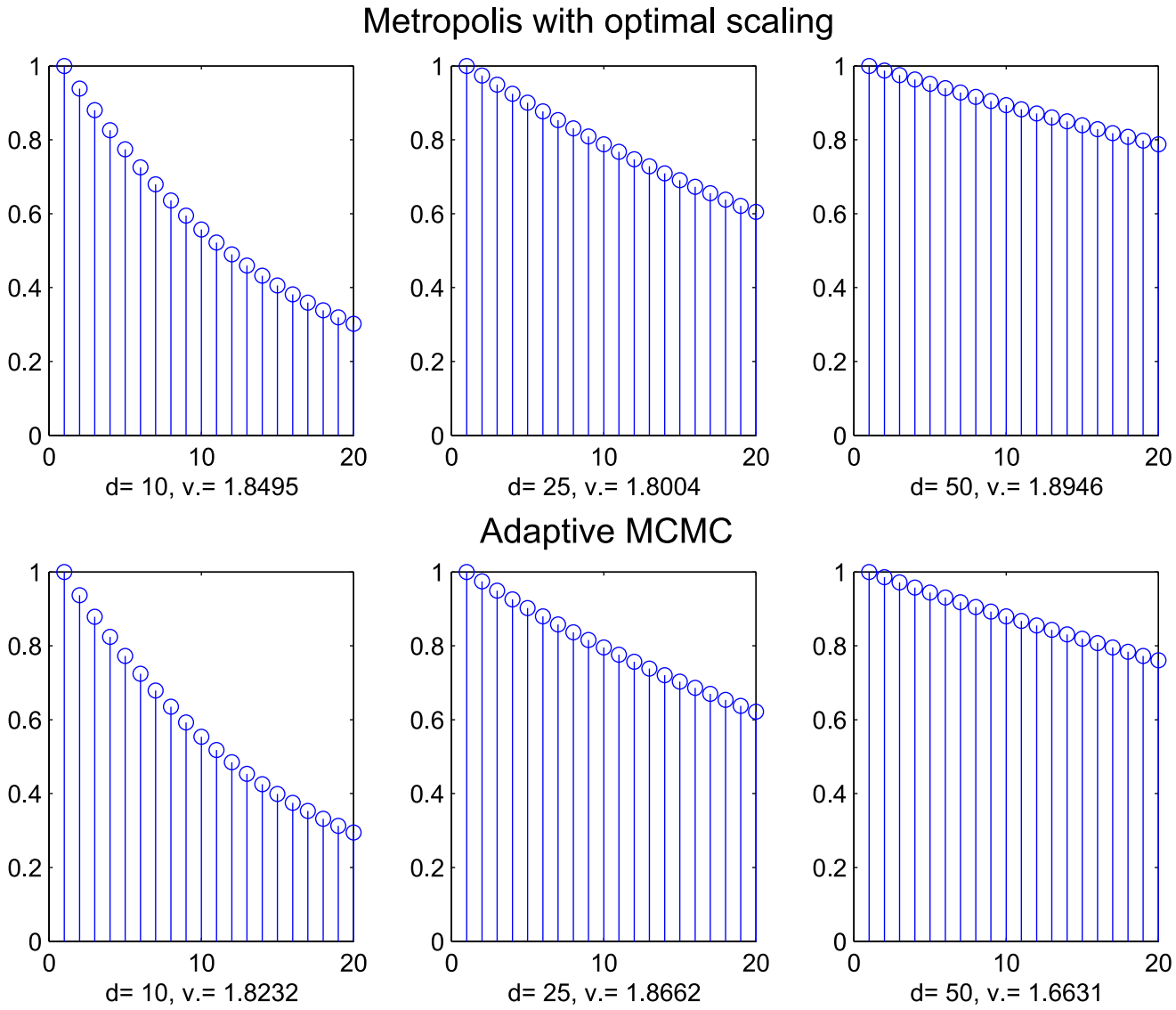
$$X_{k+1} = \begin{cases} Y_{k+1} & \textbf{with prob. } \alpha(X_k, Y_{k+1}) \\ X_k & \textbf{otherwise} \end{cases}$$

- Update the scaling factor

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

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

- A better form of this algorithm is discussed in **?**


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

Adaptive MCMC with multidimensional scaling (Haario, Saksman, Tamminen, 1999, 2001), Vihola (2010)

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

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

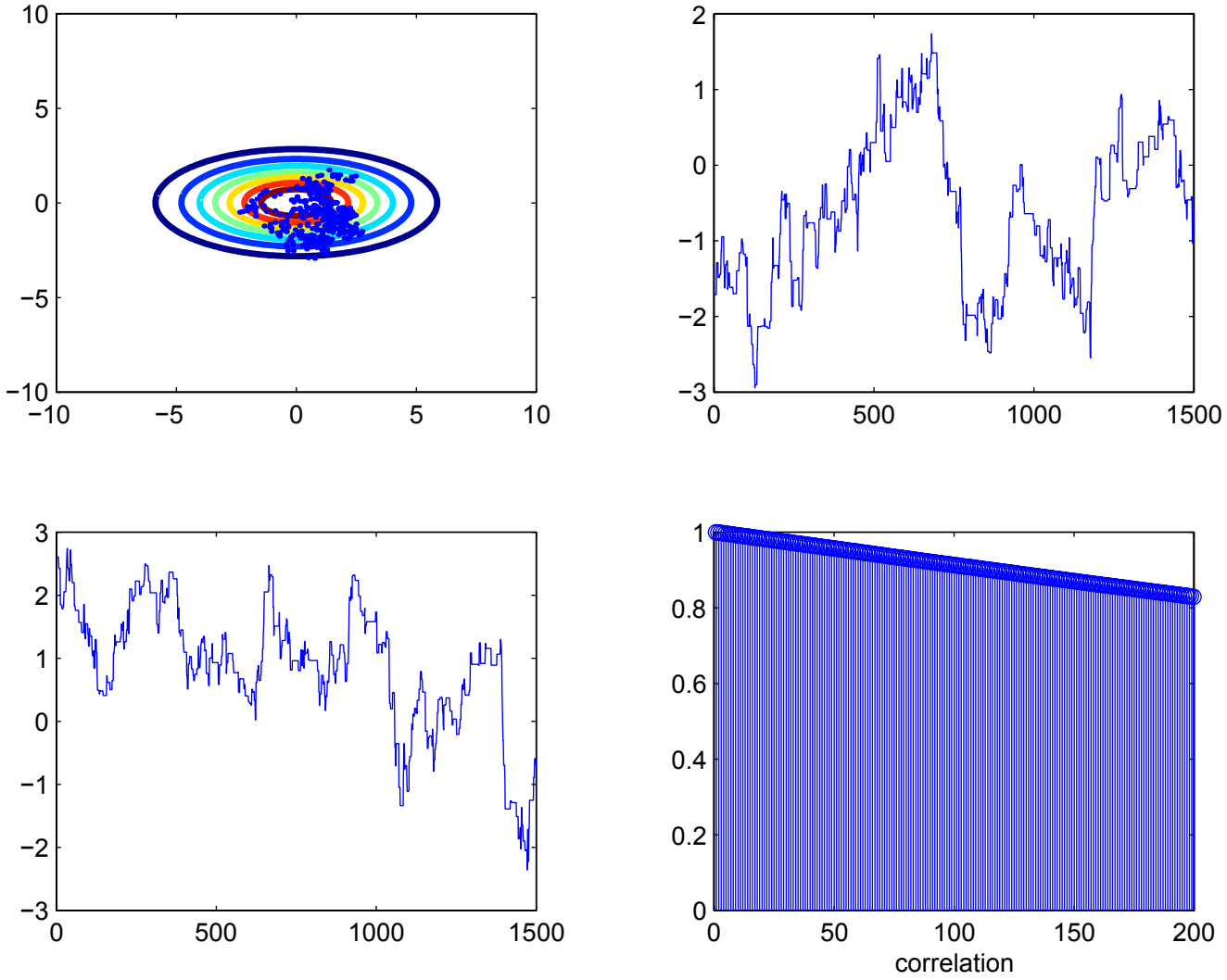


Figure: $d = 12$, $\pi \sim \mathcal{N}(0, \Gamma)$, $\mathrm{cond}(\Gamma) \approx 100$, $q \sim \mathcal{N}(0, (2.32^2/d)\,\mathrm{I})$

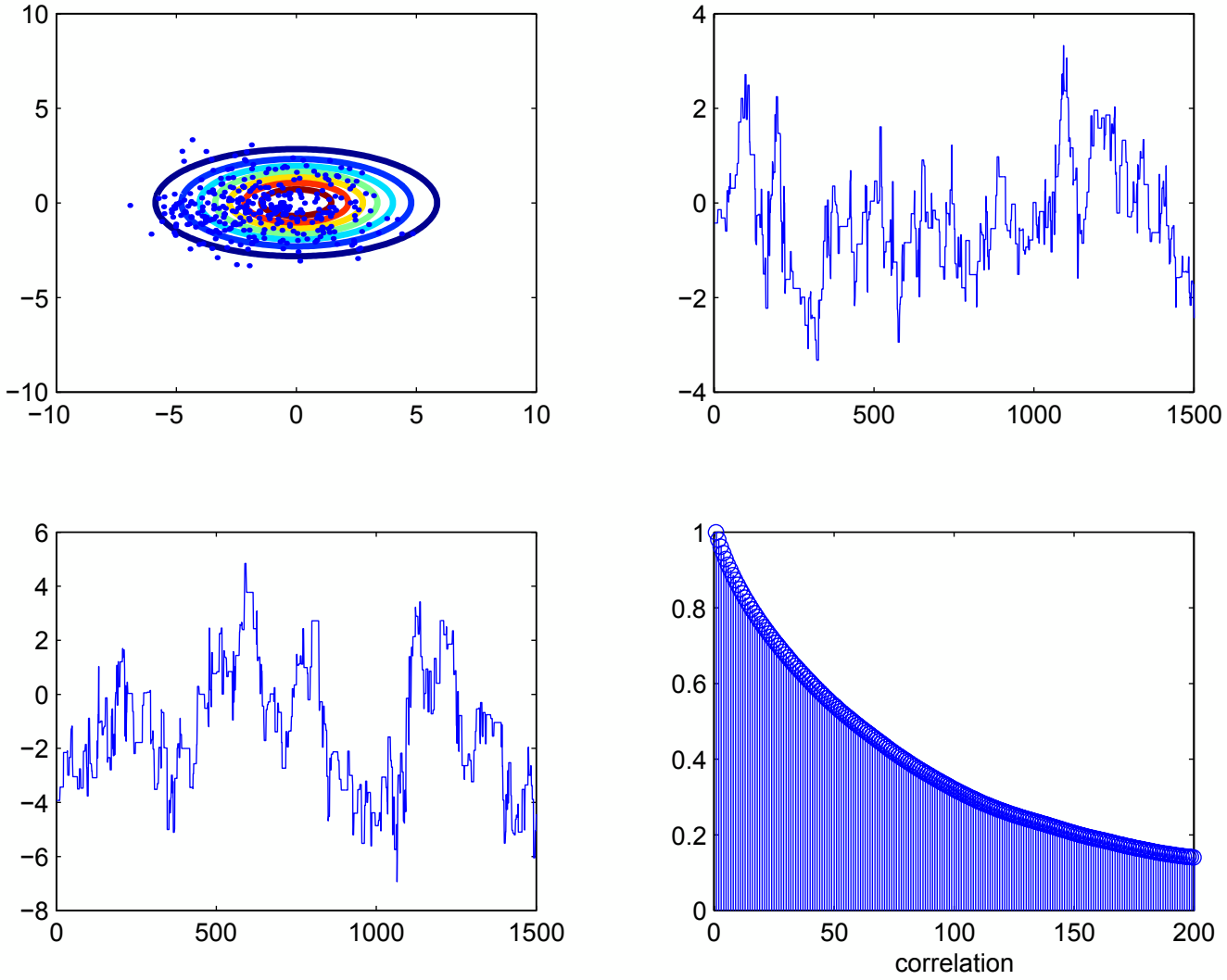


Figure: $d = 12$, $\pi \sim \mathcal{N}(0, \Gamma)$, $\mathrm{cond}(\Gamma) \approx 100$, $q \sim \mathcal{N}(0, (2.32^2/d)\,\Gamma)$

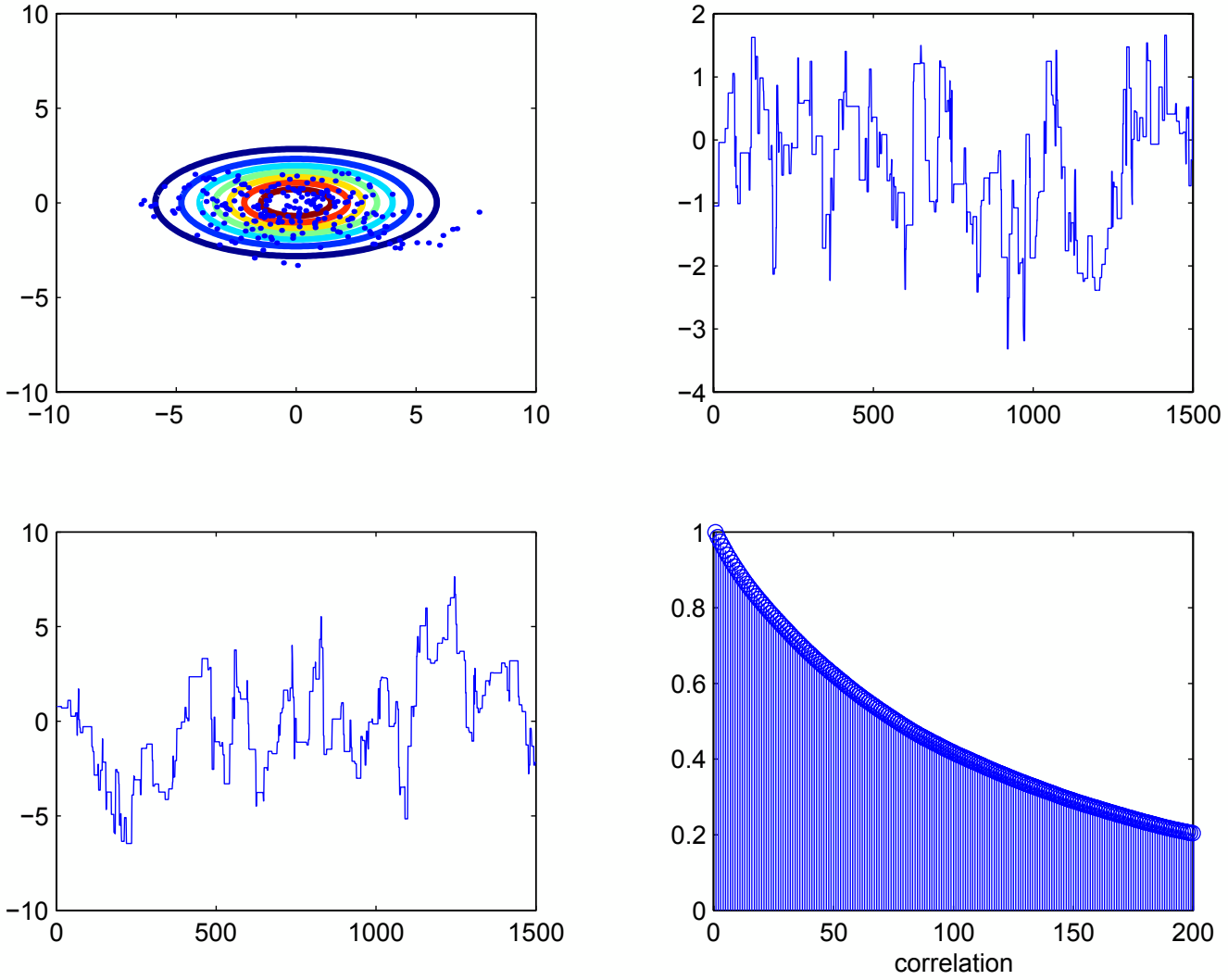


Figure: $d = 12$, $\pi \sim \mathcal{N}(0, \Gamma)$, $\mathrm{cond}(\Gamma) \approx 100$, $q \sim \mathcal{N}(0, \sigma_k \Gamma_k)$, with adaptive multidimensional scaling

Tempering

- MCMC generally run into troubles when the distribution is multimodal.
- Discovering modes is to 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 (introduced by Kou, Zhou, Wong, 2008) 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

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

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 $\alpha_{n,k}(X_n^{(k)}, X_\star^{(k-1)})$

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

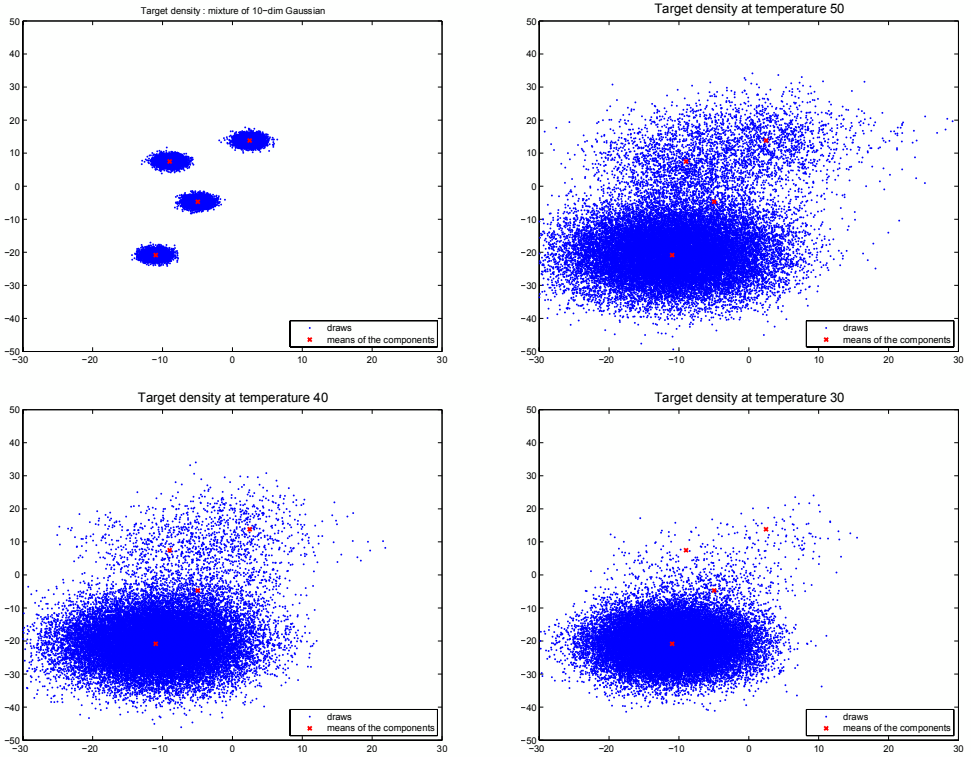
where $\theta_n^{(k-1)}(\mathrm{d}y) = \frac{1}{n+1}\sum_{\ell=1}^n \delta_{X_\ell^{(k-1)}}(\mathrm{d}y)$ and

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

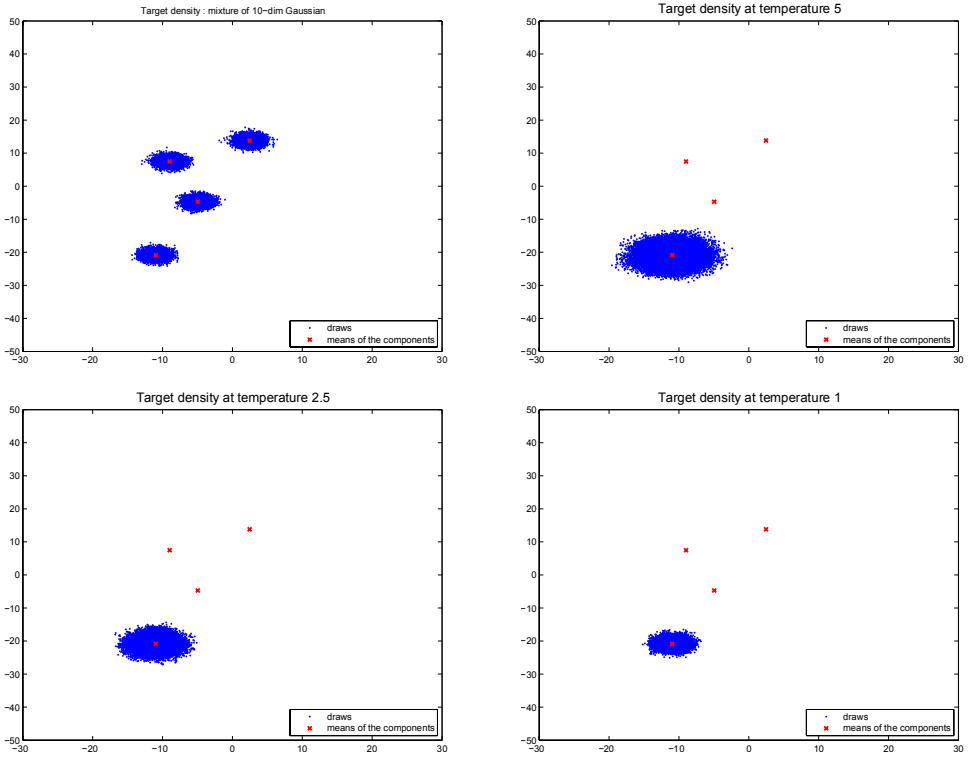
An example

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).

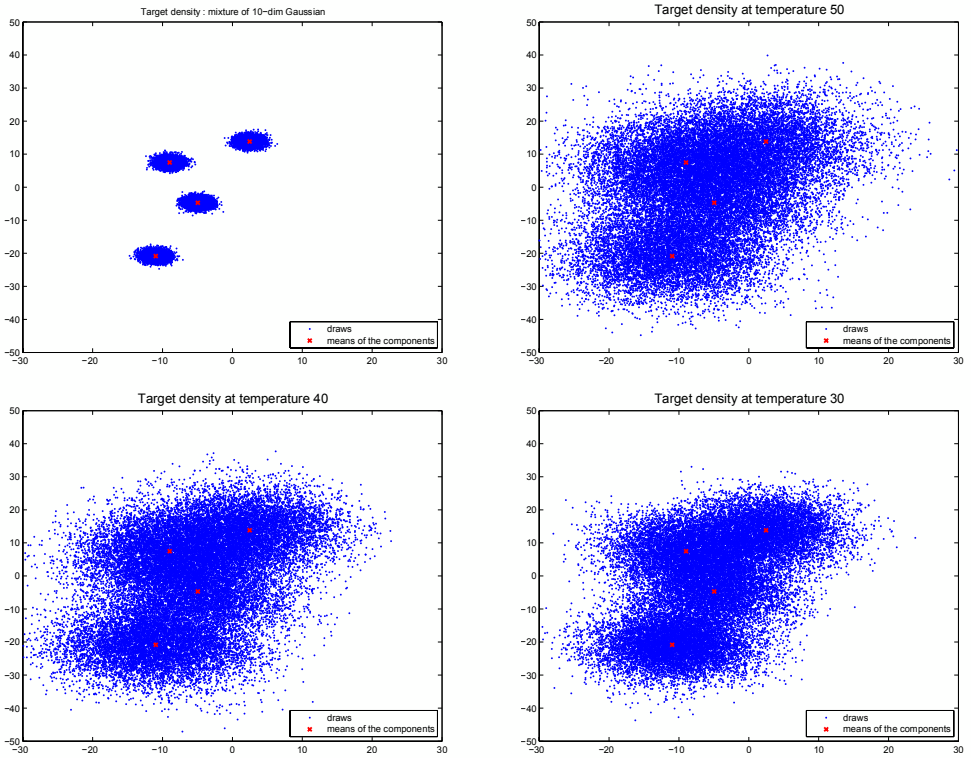
Interactions: 0.001 %



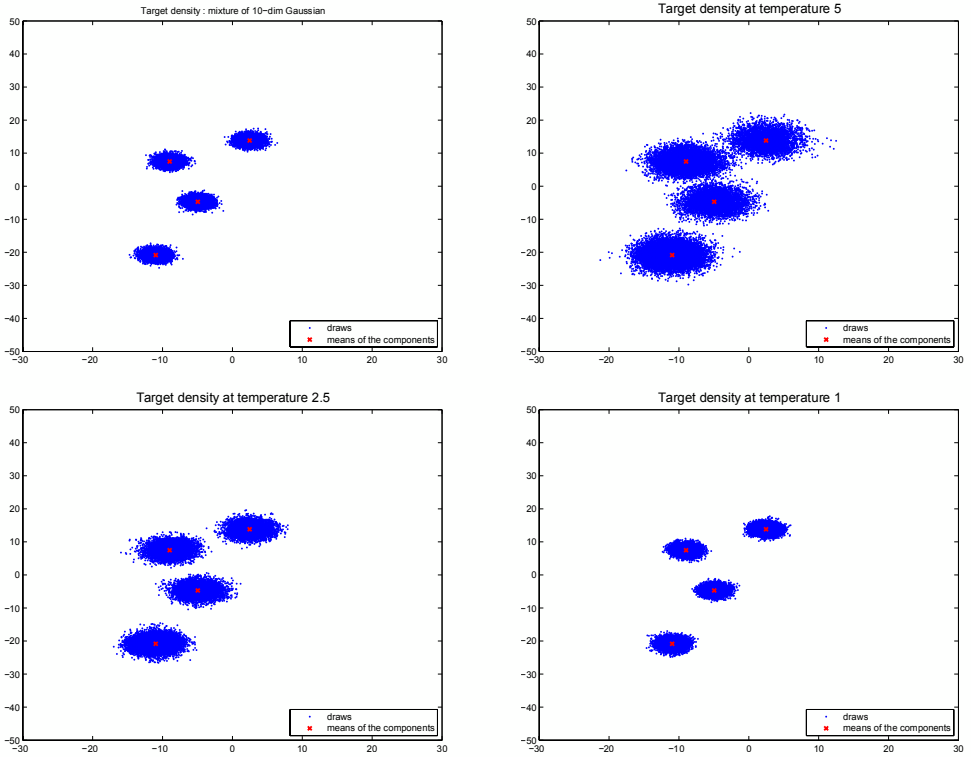
Interactions: 0.001 %



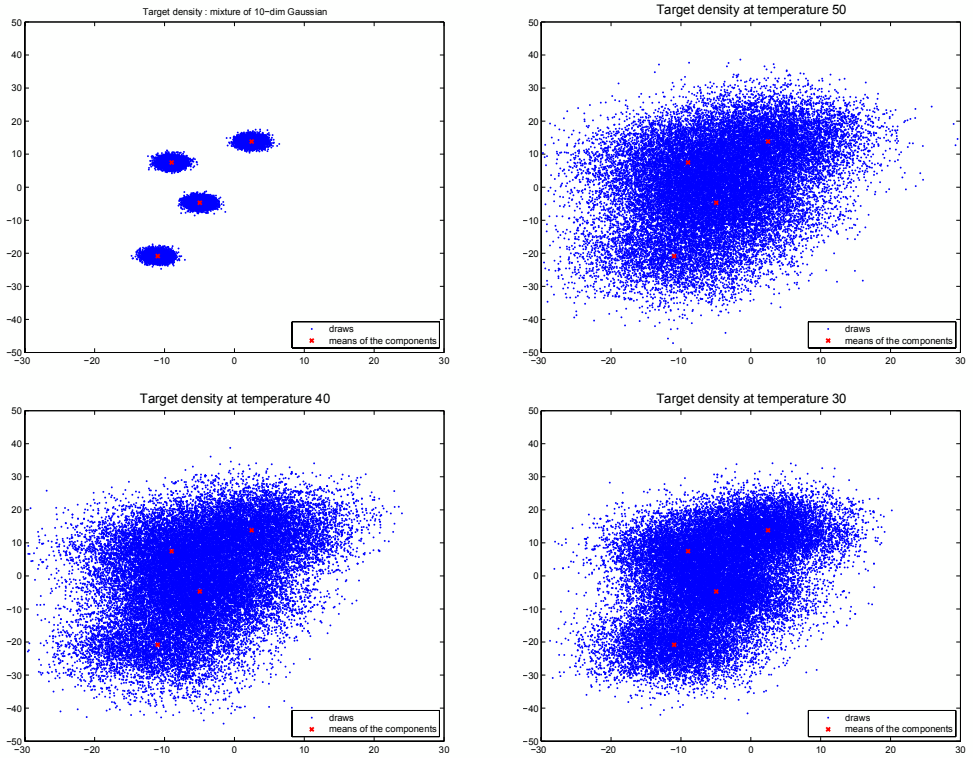
Interactions: 5 %



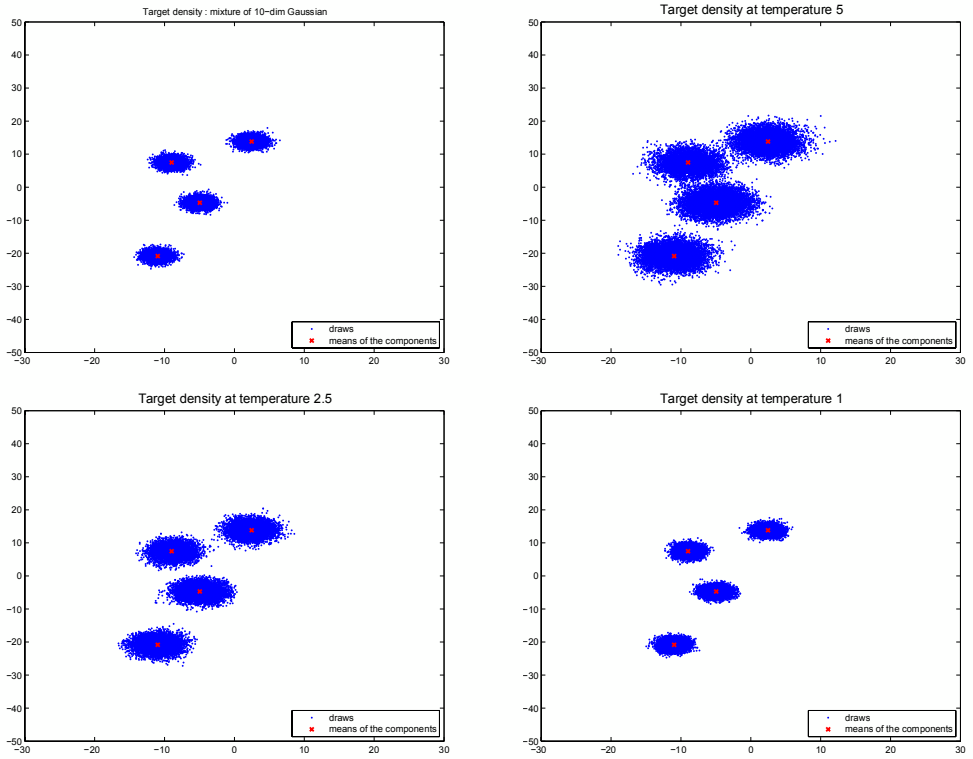
Interactions: 5 %



Interactions: 20 %



Interactions: 20 %



Definitions and General Assumptions

- Let $(\Theta, \mathcal{T})$ be a measurable space and $(\mathsf{X}, \mathcal{X})$ be 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 unique stationary 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 $(\Omega, \mathcal{F}, \mathcal{F}_n, \mathbb{P})$ such that, for each n, (X_n, θ_n) is $\mathcal{F}_n$-measurable 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.
  3. **Central Limit Theorems** for additive functionals: $n^{-1/2} \sum_{k=1}^{n} \{f(X_k) - \pi(f)\} \Rightarrow_d \mathcal{N}(0, \sigma^2)$ [not today !]
- **Problem:** $\{X_k\}$ is **not** a Markov Chain.

A quick survey of the literature: Adaptive MCMC

I hope that I did not have forgotten anyone in the room !

1. Haario, Saksman, Tamminen (1999), (2001): analysis of the Adaptive Metropolis under some strong assumptions (use mixingales approach)
2. Andrieu, Moulines (2006): analysis of a general class of adaptive algorithms where the parameter is adapted using Stochastic Approximation
3. Roberts and Rosenthal (2007): more general algorithms (**diminishing adaptations** and the **containment conditions**). Some of these conditions are found to be close to necessary; see Bai, Roberts, Rosenthal (2010)
4. Atchadé and Fort (2009): adaptation to sub-geometric convergence
5. Saksman & Vihola (2010): a version of Adaptive Metropolis without strange looking conditions !

The level-0 asymptotic theory is almost complete... the challenge is now to develop some finite horizon results, hopefully showing that adaptation is useful even if the number of iterations is finite.

A quick survey of the literature: Interacting MCMC

1. Del Moral and Miclo (2004): self-interacting chains under uniform ergodicity conditions (a curiosity)
2. Andrieu, Del Moral, Doucet, Jasra (2006,2007,2010): law of large numbers for a "two-level" process under geometric ergodicity conditions (the proof if Kou, Zhou, Wong (2006) is flawed)
3. Del Moral and Doucet (2009), Bercu, Del Moral and Doucet (2009): general interacting algorithms, for uniformly ergodic chain. Some conditions are likely to be very hard to check (and are not satisfied for "classical" algorithms)
4. Atchade (2009): an attempt to solve the general problem. Some gaps in the proof.

Error decomposition

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

Error decomposition

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

↪ [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 $\{r_n, n \geq 0\}$ such that $\limsup_{n\to\infty} r_n/n = 0$ and

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

Error decomposition

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

$\hookrightarrow$ [B] (Diminishing adaptation)
For any $\epsilon > 0$,

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

Error decomposition

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

$\hookrightarrow$ [C] Convergence of the invariant distributions

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

Result [Fort et al. 2010]

Theorem

Assume (A)-(B)-(C). Then, $\lim_n \mathbb{E}\left[f(X_n)\right] = \pi(f)$.

Condition (A) is easily checked if the kernel is geometrically ergodic

$$P_\theta V \leq \lambda_\theta V + b_\theta \ ,$$

$$P_\theta(x,\cdot) \geq \delta_\theta \ \nu_\theta(\cdot) \ \mathbb{1}_{\{V \leq c_\theta\}}(x) \qquad c_\theta \stackrel{\text{def}}{=} 2b_\theta(1-\lambda_\theta)^{-1} - 1 \ .$$

In such case

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

and there exists universal constants C and γ s.t.

$$L_\theta \stackrel{\text{def}}{=} C_\theta \vee (1-\rho_\theta)^{-1} \leq C\left\{b_\theta \vee \delta_\theta^{-1} \vee (1-\lambda_\theta)^{-1}\right\}^\gamma \ .$$

Comparison with [Roberts & Rosenthal, 2007]

1. Weaken the **containment condition** and the **diminishing adaptation condition** of [Roberts & Rosenthal, 2007]. For example (A) cover situations where the transition kernels P_θ are geometrically ergodic but **not necessarily** uniformly-in-θ.

$$\sup_{f,|f|\leq V} |P_\theta^n f(x) - \pi_\theta(f)| \leq C_\theta \ \rho_\theta^n \ V(x)$$

   Nevertheless, it is required to have an explicit control of ergodicity s.t. $C_{\theta_n} \vee (1-\rho_{\theta_n})^{-1}$ does not "explode too quickly".
2. π_θ can depend upon θ provided we are able to prove that $\pi_{\theta_n}(f)$ converges to $\pi(f)$.
3. the analysis of the unconstrained adaptive MCMC (for which the containment condition seems to be difficult to check directly) and of a simplified version of the interacting tempering is given in (Fort, et al., 2010)

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 second term, required to prove that $\pi_{\theta_n}(f) \xrightarrow{\text{a.s.}} \pi(f)$

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, replace $f - \pi_{\theta_{k-1}}(f)$ by $\hat{f}_{\theta_{k-1}} - P_{\theta_{k-1}}\hat{f}_{\theta_{k-1}}$ where $\hat{f}_\theta$ is the solution of the **Poisson Equation**

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

Decomposition of the error

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

$$=\frac{1}{n}\sum_{k=1}^{n-1}\{\hat{f}_{\theta_{k-1}}(X_k)-P_{\theta_{k-1}}\hat{f}_{\theta_{k-1}}(X_{k-1})\}+R_n^{(1)}+R_n^{(2)}$$

where $R_n^{(1)}$ and $R_n^{(2)}$ are remainder terms

$$R_n^{(1)} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{k=1}^{n-1}\{P_{\theta_k}\hat{f}_{\theta_k}(X_k)-P_{\theta_{k-1}}\hat{f}_{\theta_{k-1}}(X_k)\}\ ,$$

$$R_n^{(2)} \stackrel{\text{def}}{=} \frac{1}{n}P_{\theta_0}\hat{f}_{\theta_0}(X_0)-\frac{1}{n}P_{\theta_{n-1}}\hat{f}_{\theta_{n-1}}(X_{n-1})\ .$$

Decomposition of the error

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

$$= \frac{1}{n} \sum_{k=1}^{n-1} \{ \hat{f}_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} \hat{f}_{\theta_{k-1}}(X_{k-1}) \} + R_n^{(1)} + R_n^{(2)}$$

↪ [B] Chow's Martingale Cvge Theorem: for some $\alpha > 1$,

$$\sum_k \frac{1}{k^\alpha} \, \mathbb{E} \left[|\hat{f}_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} \hat{f}_{\theta_{k-1}}(X_{k-1})|^\alpha \, \middle| \, \mathcal{F}_{k-1} \right] < +\infty \quad \text{a.s.}$$

Decomposition of the error

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

$$= \frac{1}{n} \sum_{k=1}^{n-1} \{ \hat{f}_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} \hat{f}_{\theta_{k-1}}(X_{k-1}) \} + R_n^{(1)} + R_n^{(2)}$$

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

Theorem

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* $\alpha > 1$

$$\sum_k k^{-\alpha} \ (L_{\theta_k})^{2\alpha} \ P_{\theta_k} V^\alpha(X_k) < +\infty \quad \text{a.s.}$$

with $L_\theta = C_\theta \vee (1-\rho_\theta)^{-1}$.

C. *(Strengthened diminishing adaptation)*

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

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

Convergence of the stationary distribution

Theorem

A. *(Ergodicity of the transition kernels)*

B. X *is Polish*

C. *(Uniform Feller condition) 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, $\pi_{\theta_n}(f) \xrightarrow{\text{a.s.}} \pi_{\theta_\star}(f)$.

Convergence of the stationary distribution

Theorem (Extended Varadarajan Theorem)

Let (U, d) *be a metric space equipped with its Borel* σ*-field* $\mathcal{B}(\mathsf{U})$*. Let* $(\Omega, \mathcal{A}, \mathbb{P})$ *be a probability space,* μ *be a distribution on* $(\mathsf{U}, \mathcal{B}(\mathsf{U}))$ *and* $\{K_n, n \geq 0\}$ *be a family of Markov transition kernels* $K_n : \Omega \times \mathcal{B}(\mathsf{U}) \to [0, 1]$*. Assume that, for any* $f \in \mathsf{C}_b(\mathsf{U}, d)$

$$\Omega_f \stackrel{\text{def}}{=} \left\{\omega \in \Omega : \limsup_{n\to\infty} |K_n(\omega, f) - \mu(f)| = 0\right\} ,$$

is a $\mathbb{P}$*-full set. Then*

$$\left\{\omega \in \Omega : \forall f \in \mathsf{C}_b(\mathsf{U}, d) \qquad \limsup_{n\to\infty} |K_n(\omega, f) - \mu(f)| = 0\right\} ,$$

is a $\mathbb{P}$*-full set.*

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

- the target density π is *lighter than exponential*
- the eigenvalues of the covariance matrix of the proposal is lower bounded by $\kappa > 0$ (loading factor).

1. **Ergodicity**: $\lim_n \sup_{f,|f|_\infty \leq 1} \mathbb{E}\left[f(X_n)\right] = \pi(f)$. see also (Bai et al., 2010)
2. **Strong law of large numbers**: for any function f such that $|f(x)| \leq \pi^{-s}(x)$, $s \in (0,1)$. most of the arguments are adapted from (Saksman & Vihola, 2009)!, used as a stress test !

Convergence of the Interacting Tempering Algorithm

- target density π is **lighter than exponential**
- Any number of stages, no restriction on the probability of swap $\epsilon \in (0,1)$, the temperature schedules, etc..
- the "local" 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)

Improved versions of the algorithms (with adaptive energy rings) have also been considered.