# A Partial Ordering for Inhomogeneous Markov Chains : Motivations and Applications to several MCMC algorithms...


Florian Maire*, UCD

joint work with : Randal Douc (Telecom SudParis, Évry, France)
Jimmy Olsson (KTH Institute of Technology, Stockholm, Sweden)

*while at ONERA & Telecom SudParis


Working Group on Statistical Learning, 5th of February 2014

# Outlines

# Outlines

1. Motivations & main Problematic
2. A new Theorem for Markov chains comparaison
3. Applications to some MCMC algorithms

# A toy example to start

Consider the joint probability distribution whose density function is defined on $(\{1,\ldots,4\},\mathbb{R}^2)$ by:

$$\pi(i,x)=\frac{1}{4}g_i(x)\ ,$$

where $\{g_i\,,\,i\in(1,4)\}$ is the Gaussian density function with mean

$$\mu_1=\begin{pmatrix}1\\1\end{pmatrix},\mu_2=\begin{pmatrix}-1\\1\end{pmatrix},\mu_3=\begin{pmatrix}-1\\-1\end{pmatrix},\mu_4=\begin{pmatrix}1\\-1\end{pmatrix}$$

and covariance matrix $\Sigma=\sigma^2\mathrm{Id}_2$.

$\Rightarrow$ We want to sample $(I,X)\sim\pi$
(and imagine no exact sampling is available)

# Inefficient Gibbs sampler... (1/2)

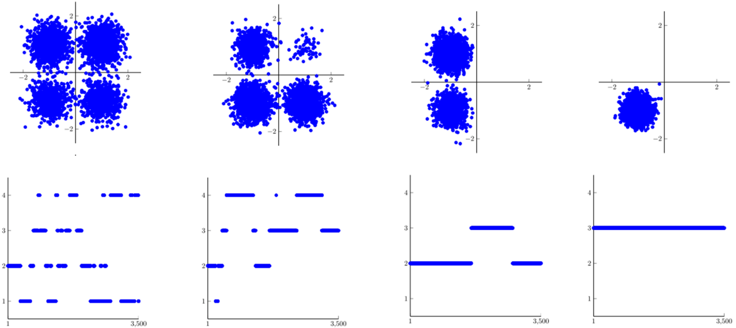
Gibbs sampler (1984): transition $(I_n = i, X_n = x) \to (I_{n+1}, X_{n+1})$ writes

(i) $X_{n+1} \mid I_n = i \sim g_i$ ,

(ii) $I_{n+1} = i' \mid X_{n+1} = x' \propto g_{i'}(x')$ .

Table: Illustration of the Gibbs Markov chain under different $\sigma^2$

$\sigma^2 = 0.125$ | $\sigma^2 = 0.1$ | $\sigma^2 = 0.075$ | $\sigma^2 = 0.005$



# Inefficient Gibbs sampler... (2/2)

Table: Empirical model transition probability obtained by the Gibbs sampler with different $\sigma$

| $\sigma^2$ | 0.125 | 0.1 | 0.075 | 0.005 |
|---|---|---|---|---|
| $\hat{\mathbb{P}}[I_{n+1} \neq I_n]\,(10^{-6})$ | 2300 | 860 | 85 | 1.2 |

- The well known Gibbs trapping state problem:

$$x \sim g_i \implies x \text{ fits model } i \implies \mathbb{P}[i \to j \neq i \mid x] \ll 1\,.$$

This problem is all the more important when models are distinct *i.e* when $\{\pi(\,\cdot \mid i), i \in (1,4)\}$ are class informative

# Another argument

- Formalism: $Z = (I, X)$ and $\pi$ defined on $(\mathsf{Z}, \mathcal{Z})$,
- Previous example is a particular case of the more general model where the state space writes

$$\mathsf{Z} = \left\{ i \in \mathsf{I},\, X \in \mathsf{X}^{(i)} \right\} ,$$

- A Gibbs sampler **cannot simulate** a Markov chain on $(\mathsf{Z}, \mathcal{Z})$
  $\Rightarrow$ the Gibbs scheme would allow samples $(i, x \in X^{(j)}) \notin \mathsf{Z}$ ,
- From now on, suppose all the parameters live in the same space *i.e*

$$\mathsf{Z} = \left\{ i \in \mathsf{I},\, x \in \mathsf{X} \right\} .$$

# A more appropriate sampler

- Suppose a mixture of $C$ models *i.e* $I \in \mathsf{I}$ ($\mathsf{I} = \{1, \ldots, C\}$),
- Carlin & Chib (1995) propose the *extended* target distribution

$$\tilde{\pi}(i, x^{(1)}, \ldots, x^{(C)}) = \pi(i, x^{(i)}) \prod_{j \neq i} \zeta_j(x^{(j)}) \, ,$$

where $\{\zeta_i, \ i \in \mathsf{I}\}$ are "samplable" probability distributions refered to as *pseudo-priors*.

- Note that:

$$\int \cdots \int \tilde{\pi}(i, \mathrm{d}x^{(1)}, \ldots, \mathrm{d}x^{(i-1)}, x^{(i)}, \mathrm{d}x^{(i+1)}, \ldots, \mathrm{d}x^{(C)}) = \pi(i, x^{(i)}) \, ,$$

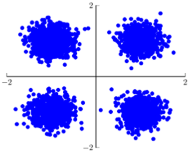
- and thus

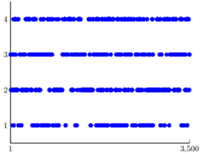
$$(I, X^{(1)}, \ldots, X^{(C)}) \sim \tilde{\pi} \Longrightarrow (I, X^{(I)}) \sim \pi \, .$$

# Carlin & Chib

- The Carlin & Chib sampler is a Gibbs on the data-augmented state-space $\mathsf{I} \times \underbrace{\mathsf{X} \times \cdots \times \mathsf{X}}_{C}$
- Given $(I_n, X_n^{(1)}, \ldots, X_n^{(C)}) = (i, x^{(1)}, \ldots, x^{(C)})$, the transition writes
  (i) $X_{n+1}^{(i)} \sim \pi(\,\cdot\,|\,i)$,
  (ii) $\forall j \neq i\,, \quad X_{n+1}^{(j)} \sim \zeta_j$,
  (iii) draw $I_{n+1} = i'$ with proba. $\propto \tilde{\pi}(i', X_{n+1}^{(1)}, \ldots, X_{n+1}^{(C)})$.

Table: Marginal sequence $\{(I_n, X_n^{(I_n)}),\, n \in \mathbb{N}\}$ when $\sigma^2 = 0.005$



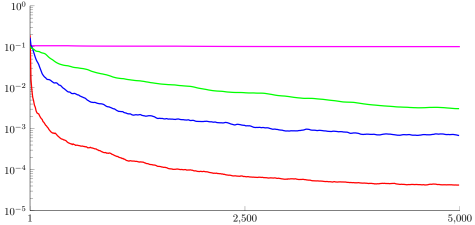



# Influence of the *pseudo-priors*

- We would like to have $\zeta_i \approx \pi(\,\cdot\,|\,i)$
- Marginal probabilities of the class after 5000 MCMC iterations

| classes | | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Gibbs | | 0 | 0 | 1 | 0 |
| CC | $\zeta_j = g_j$ | 0.26 | 0.24 | 0.25 | 0.25 |
| CC | $\zeta_j = \mathcal{N}(0,1)$ | 0.24 | 0.27 | 0.23 | 0.26 |
| CC | $\zeta_j = \mathcal{N}(0,0.2)$ | 0.44 | 0.17 | 0.25 | 0.14 |

- Evolution of the empirical variance of $\mathbb{P}[I = 1]$ throughout MCMC



# Theoretic considerations

Is there any theoretic argument behind the (obvious) link between

(i) *the ability of the Markov chain to switch models*

(ii) *the MCMC asymptotic variance?*

Formalizing the *ability to switch between models* leads to the **off-diagonal** ordering:

- Let $P_0$ and $P_1$ two Markov kernels on some general state space $(\mathsf{Z}, \mathcal{Z})$
- $P_1$ dominates $P_0$ in the off-diagonal sense if $\forall\, A \in \mathcal{Z}$

$$P_1(z, A \backslash \{z\}) \geq P_0(z, A \backslash \{z\}), \quad \pi\text{-a.e}\,.$$

(we note $P_1 \succeq P_0$)

# Tierney's Theorem

Tierney (1994) Theorem (extending Peskun's (1973)) state that :

Under (A1) and (A2)

- (A1) $P_0$ and $P_1$ are $\pi$-reversible kernels *i.e* for $i \in \{0, 1\}$:

$$\forall (A, B) \in (\mathcal{Z} \times \mathcal{Z}) \,, \qquad \int_A \pi(\mathrm{d}z) P_i(z, B) = \int_B \pi(\mathrm{d}z) P_i(z, A) \,,$$

- (A2) $P_1 \succeq P_0$ ,

Then, **for all** $f \in \mathcal{L}^2(\pi)$

$$v(f, P_1) \leq v(f, P_0) \,,$$

where for $i \in \{0, 1\}$,

$$v(f, P_i) := \lim_{n \to \infty} \frac{1}{n} \mathrm{Var}\left[\sum_{k=1}^{n} f(X_k^{(i)})\right] \,, \quad X_k^{(i)} \sim P_i^{(k)}(z_0, \cdot\,) \,.$$

# Limitations

- Some popular kernels may not have feature the *off-diagonal* ordering
- (A1) ($\pi$-reversibility) is a strong assumption ...
- ...and is not verified by either Gibbs or Carlin & Chib sampler
- To obtain a $\pi$-reversible Markov chain, Gibbs sampler (and Carlin & Chib) should be rewritten as :

$$\begin{pmatrix} I_n = i \\ X_n = x \end{pmatrix} \to \begin{pmatrix} I_{n+1} \sim \pi(\,\cdot\,|\,x) \\ X_{n+1} \sim \delta_{\{x\}}(\,\cdot\,) \end{pmatrix} \to \begin{pmatrix} I_{n+2} \sim \delta_{\{i'\}}(\,\cdot\,) \\ X_{n+2} \sim \pi(\,\cdot\,|\,i') \end{pmatrix} \to \cdots$$

- that is **Inhomogeneous** Markov chain

$$Z_n \xrightarrow{P} Z_{n+1} \xrightarrow{Q} Z_{n+2} \xrightarrow{P} Z_{n+3} \xrightarrow{Q} \cdots$$

which is not in Tierney's Theorem scope.

# Question

Is there any way to extend Tierney's Theorem to cover the inhomogeneous Markov chain study?

Tierney's proof essentially relies on

(i) the following expression of the variance:

$$\frac{1}{n}\mathrm{Var}\left[\sum_{k=1}^{n} f(Z_k)\right] = \|f\|^2 + \frac{2}{n}\sum_{k=1}^{n}(n-k)\left\langle f, P^k f\right\rangle \ ,$$

(ii) a spectral decomposition Theorem for self-adjoint operators:

$$\forall\, n \geq 0, \quad \langle f, P^n f\rangle = \int z^n \mu_{f,P}(\mathrm{d}z) \ .$$

A similar Proof cannot be derived for inhomogeneous chains.

# Outlines

# Our Main result

Under (A1') and (A2')

- (A1') for all $i \in \{0, 1\}$, $P_i$ and $Q_i$ are $\pi$-reversible kernels
- (A2') $P_1 \succeq P_0$ and $Q_1 \succeq Q_0$

Then, for all $f \in \mathcal{L}^2(\pi)$ such that

$$\sum_{k=1}^{\infty} \left( |\mathrm{Cov}(f(Z_0^{(i)}), f(Z_k^{(i)}))| + |\mathrm{Cov}(f(Z_1^{(i)}), f(Z_{k+1}^{(i)}))| \right) < \infty \,, \qquad (1)$$

we have

$$v(f, P_1, Q_1) \leq v(f, P_0, Q_0) \,.$$

# Sketch of the proof

$\Rightarrow$ Revisiting Tierney's proof without spectral decomposition Theorem

(1) Under assumptions (A1) and (A2) and for $f$ such that (1)

$$v(f,P) = \|f\|^2 + 2\sum_{n=0}^{\infty} \underbrace{\mathrm{Cov}(f(X_1), f(X_n))}_{\langle f, P^n f\rangle} ,$$

(2) Define $\begin{cases} \forall\, \alpha \in (0,1) & P_\alpha = (1-\alpha)P_0 + \alpha P_1 , \\ \forall\, \lambda \in (0,1) & w_\lambda(f, P_\alpha) = \sum_{n=1}^{\infty} \lambda^n \langle f, P_\alpha^n f\rangle , \end{cases}$

(3) We show that $\forall\, \lambda \in (0,1)$, $\alpha \to w_\lambda(f, P_\alpha)$ is decreasing over $(0,1)$.

(4) Proof completed by a Dominated Convergence Theorem $\lambda \to 1$ :

$$\langle f, P_1^n f\rangle \leq \langle f, P_0^n f\rangle .$$

$\Rightarrow$ This proof is compatible with inhomogeneous Markov chain.

# A significant Corollary

- Imagine $Z = (X, U)$ where $X$ is the **variable of interest** and $U$ **some auxiliary data**
- In many situation the transition kernel $K_i$ isn't $\pi$-reversible

$$Z_n^{(i)} \xrightarrow{K_i} Z_{n+1}^{(i)} \xrightarrow{K_i} Z_{n+2}^{(i)} \cdots$$

and thus $K_1 \succeq K_0 \not\Rightarrow v(f, K_1) \leq v(f, K_0)$ (with Tierney Theorem)

- Possibility to "force" $\pi$-reversibility by artificially introducing a *freezing* step

$$\tilde{Z}_n = \begin{pmatrix} \tilde{X}_n^{(i)} \\ \tilde{U}_n^{(i)} \end{pmatrix} \xrightarrow{P_i} \begin{pmatrix} \tilde{X}_{n+1}^{(i)} \\ \tilde{U}_{n+1}^{(i)} = \tilde{U}_n^{(i)} \end{pmatrix} \xrightarrow{Q_i} \begin{pmatrix} \tilde{X}_{n+2}^{(i)} = \tilde{X}_{n+1}^{(i)} \\ \tilde{U}_{n+2}^{(i)} \end{pmatrix} \xrightarrow{P_i} \cdots$$

- Note that $\{\tilde{Z}_{2n}^{(i)},\, n \in \mathbb{N}\} = \{Z_n^{(i)},\, n \in \mathbb{N}\}$ and our Theorem leads to $v(f, K_1) \leq v(f, K_0)$

# Outlines

# The mixture Model problem

Rewriting the Gibbs and the Carlin and Chib samplers:

$$P_{\mathtt{G}}\begin{cases} I_n^{(\mathtt{G})} \sim \pi(\,\cdot\,|\,x) \\ X_n^{(\mathtt{G})} \sim \delta_x(\,\cdot\,) \end{cases} \qquad\qquad Q_{\mathtt{G}}\begin{cases} I_{n+1}^{(\mathtt{G})} \sim \delta_{i'}(\,\cdot\,) \\ X_{n+1}^{(\mathtt{G})} \sim \pi(\,\cdot\,|\,i') \end{cases}$$

$$\begin{tikzcd}
 & Z_{2n+1}^{(\mathtt{G})} = \begin{pmatrix} I_n^{(\mathtt{G})} = i' \\ X_n^{(\mathtt{G})} = x \end{pmatrix} \arrow[dr] & \\
Z_{2n}^{(i)} = \begin{pmatrix} I_n^{(i)} = i \\ X_n^{(i)} = x \end{pmatrix} \arrow[ur] \arrow[dr] & & Z_{2n+2}^{(i)} = \begin{pmatrix} I_{k+1}^{(i)} = i' \\ X_{k+1}^{(i)} = x' \end{pmatrix} \\
 & Z_{2n+1}^{(\mathtt{CC})} = \begin{pmatrix} I_n^{(\mathtt{CC})} = i' \\ X_n^{(\mathtt{CC})} = u_{i'} \end{pmatrix} \arrow[ur] &
\end{tikzcd}$$

$$P_{\mathtt{CC}}\begin{cases} \forall\, i \neq j \;\; U_i \sim \zeta_i(\,\cdot\,) \\ U_j \sim \delta_x(\,\cdot\,) \\ I_n^{(\mathtt{CC})} \sim \pi(\,\cdot\,|\,u_1,\ldots,u_C) \end{cases} \qquad\qquad Q_{\mathtt{CC}}\begin{cases} I_{n+1}^{(\mathtt{CC})} \sim \delta_{i'}(\,\cdot\,) \\ X_{n+1}^{(\mathtt{CC})} \sim \pi(\,\cdot\,|\,i') \end{cases}$$

(i) $P_{\mathtt{CC}}$ is $\pi$-reversible, (ii) $Q_{\mathtt{CC}} = Q_{\mathtt{G}}$, (iii) $P_{\mathtt{CC}} \overset{?}{\succeq} P_{\mathtt{G}} \Rightarrow v(f, \mathtt{CC}) \leq v(f, \mathtt{G})$ .

# Pseudo-Marginal Algorithms

Pseudo-Marginal (Andrieu & Robert, 2009): no exact expression of the target distribution $\pi$ *e.g*

$$\pi(x) = \int \pi(x, \mathrm{d}u)$$

for all $(x, x') \in \mathsf{X}^2$, $\pi(x)/\pi(x')$ intractable

$\Rightarrow$ Idea: simulate a Markov chain targeting

$$\tilde{\pi}(\mathrm{d}x, \mathrm{d}u) = \underbrace{\pi(\mathrm{d}x) w_u(x)}_{\hat{\pi}_u(\mathrm{d}x),\text{calculable}} \underbrace{R(x, \mathrm{d}u)}_{\text{samplable}}$$

(note that $\int \tilde{\pi}(x, \mathrm{d}u) = \pi(x)$)
For example, use Importance Sampling estimate:

$$\hat{\pi}_u(x) = \frac{1}{n} \sum_{k=1}^{n} \frac{\pi(x, u^{(k)})}{R(x, u^{(k)})}, \quad U^{(k)} \overset{i.i.d}{\sim} R(x, \cdot\,)\,.$$

# Monte Carlo within Metropolis (MCWM)

A Markov chain $\{X_n,\ n \in \mathbb{N}\}$ on $(\mathsf{X}, \mathcal{X})$: given $X_n = x$, $X_{n+1}$ is obtained as follows

(i) propose $X' \sim K(x, \cdot\,)$

(ii) simulate aux. var. for both states $X_n$ and $X'$:
$U \sim R(x, \cdot\,), U' \sim R(x', \cdot\,)$

(iii) accept $X_{n+1} = x'$ w.p

$$\hat{\alpha}(x, x', u, u') = 1 \wedge \frac{\hat{\pi}_{u'}(x')K(x', x)}{\hat{\pi}_u(x)K(x, x')}$$

MCWM is not $\pi$-reversible but targets an approximate of $\pi(x)$...
$\Rightarrow$ noisy algorithm!

# Grouped-Independence Metropolis Hastings (GIMH)

A Markov chain $\{(X_n, U_n),\ n \in \mathbb{N}\}$ targeting $\tilde{\pi}$ such that given $(X_n, U_n) = (x, u)$, $(X_{n+1}, U_{n+1})$ is obtained as follows

(i) propose $X' \sim K(x, \cdot\,)$

(ii) simulate aux. var. for the state $X'$: $U' \sim R(x', \cdot\,)$

(iii) accept $(X_{n+1}, U_{n+1}) = (x', u')$ w.p

$$\hat{\alpha}((x, u), (x', u')) = 1 \wedge \frac{\hat{\pi}_{u'}(x')K(x', x)}{\hat{\pi}_u(x)K(x, x')}$$

GIMH is **Metropolis-Hastings** algorithm $\Rightarrow$ $\tilde{\pi}$-reversible.

# Remark

- MCWM & GIMH cannot be properly compared with Tierney's Theorem
- They may be rewritten artificially as:

$$\text{MCWM:} \quad \left( X_n^{(M)} \right) \xrightarrow{P_M} \begin{pmatrix} X_n^{(M)} \\ U \end{pmatrix} \xrightarrow{Q} \left( X_{n+1}^{(M)} \right) \xrightarrow{P_M} \cdots$$

$$\text{GIMH:} \quad \begin{pmatrix} X_n^{(G)} \\ U_n^{(G)} \end{pmatrix} \xrightarrow{P_G} \begin{pmatrix} X_n^{(G)} \\ U_n^{(G)} \end{pmatrix} \xrightarrow{Q} \begin{pmatrix} X_{n+1}^{(G)} \\ U_{n+1}^{(G)} \end{pmatrix} \xrightarrow{P_G} \cdots$$

# A Random-Refreshment Pseudo Marginal algorithm

A Markov chain $\{(X_n, U_n),\ n \in \mathbb{N}\}$ targeting $\tilde{\pi}$ such that given $(X_n, U_n) = (x, u)$, $(X_{n+1}, U_{n+1})$ is obtained as follows:

(i) (a) propose a new aux. var. for state $X$; $\tilde{U} \sim R(x, \cdot\,)$
(b) refresh the aux. var. $U_n$ by $\tilde{U}$ with a certain probability $\omega_{u,\tilde{u}}$

(ii) propose $X' \sim K(x, \cdot\,)$

(iii) simulate aux. var. for the state $X'$: $U' \sim R(x', \cdot\,)$

(iv) accept $(X_{n+1}, U_{n+1}) = (x', u')$ w.p

$$\hat{\alpha}((x,u),(x',u')) = 1 \wedge \frac{\hat{\pi}_{u'}(x')K(x',x)}{\hat{\pi}_u(x)K(x,x')}$$

# Comparing GIMH & Random Refreshment

GIMH: $\begin{pmatrix} X_n^{(G)} \\ U_n^{(G)} \end{pmatrix} \xrightarrow{P_G} \begin{pmatrix} X_n^{(G)} \\ U_n^{(G)} \end{pmatrix} \xrightarrow{Q} \begin{pmatrix} X_{n+1}^{(G)} \\ U_{n+1}^{(G)} \end{pmatrix} \xrightarrow{P_G} \cdots$

Random Refreshment: $\begin{pmatrix} X_n^{(R)} \\ U_n^{(R)} \end{pmatrix} \xrightarrow{P_R} \begin{pmatrix} X_n^{(R)} \\ \tilde{U} \end{pmatrix} \xrightarrow{Q} \begin{pmatrix} X_{n+1}^{(R)} \\ U_{n+1}^{(R)} \end{pmatrix} \xrightarrow{P_G} \cdots$

Our Theorem holds and show that for any $f \in \mathcal{L}^2(\pi)$ verifying (1)

$$v(f, R) \leq v(f, G) \,.$$

# Perspectives

Our Theorem extends Tierney's and Peskun's works and allows

- to compare inhomogeneous Markov chains (by nature)...
- and even (non necessarily $\pi$-reversible) homogeneous Markov chains

Open questions remain!

- what about inhomogeneous Markov chain with $n > 2$ kernels
- possibility to find other applications such that ABC computation, and "MCMC for doubly intractable distributions", (Single Auxiliary Variable Method, Exchange Algorithm...)