

Méthodes de Monte Carlo par Chaînes de Markov adaptatives

Gersende FORT

LTCI, CNRS / TELECOM ParisTech

Méthodes MCMC adaptatives

Vers les procédures MCMC adaptatives

Paramètres d'implémentation du SRW-HM

Algorithme de Wang-Landau

Conclusion

Convergence des algorithmes MCMC adaptatifs

L'adaptation peut détruire la convergence

Conditions pour la convergence des marginales

Adaptation et Ergodicité

Résultat

En pratique, comment vérifier les conditions?

Doit-on avoir convergence du processus d'adaptation?

Conclusion

Références

Méthodes Monte Carlo par Chaînes de Markov

Etant donnée une loi cible $\pi d\nu$, il s'agit de construire une chaîne de Markov de sorte que

- $\pi d\nu$ soit la probabilité stationnaire de la chaîne.
- la chaîne vérifie des propriétés de convergence :
 - de type **ergodicité** :

$$\sup_{\{\phi, |\phi|_{\infty} \leq 1\}} \left| \mathbb{E}_x [\phi(X_n)] - \int \phi \pi d\nu \right| \rightarrow 0 \quad \forall x \in X$$

- ou de type **loi des grands nombres** :

$$\frac{1}{n} \sum_{k=1}^n \phi(X_k) \xrightarrow{\text{p.s.}} \int \phi \pi d\nu$$

- et éventuellement de type **TLC**

$$\frac{\sqrt{n}}{\sigma} \left(\frac{1}{n} \sum_{k=1}^n \phi(X_k) - \int \phi \pi d\nu \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Principalement, deux grandes familles d'algorithmes MCMC

④ algorithmes Hastings-Metropolis

- se donner un noyau de proposition $q(x,y) d\nu(y)$.
- Mise à jour $X_n \rightarrow X_{n+1}$: proposer un candidat $Y \sim q(X_n, \cdot) d\nu$ et

$$\begin{cases} X_{n+1} = Y & \text{avec probabilité } \alpha(X_n, Y) \\ X_{n+1} = X_n & \text{sinon} \end{cases}$$

où le ratio d'acceptation-rejet est donné par

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

Principalement, deux grandes familles d'algorithmes MCMC

① algorithmes Hastings-Metropolis

② algorithmes de Gibbs

- Décomposer l'espace d'état en sous-blocs $x = (x^{(1)}, \dots, x^{(d)})$
- Mise à jour $X_n \rightarrow X_{n+1}$: simuler le sous-bloc $X_{n+1}^{(i)}$ sous la loi conditionnelle du bloc i sachant les autres blocs

$$\propto \pi \left(X_{n+1}^{(1)}, \dots, X_{n+1}^{(i-1)}, y, X_n^{(i+1)}, \dots, X_n^{(d)} \right)$$

Principalement, deux grandes familles d'algorithmes MCMC

① algorithmes Hastings-Metropolis

② algorithmes de Gibbs

③ et les versions hybrides : *Metropolis within Gibbs*

dans l'algorithme de Gibbs, remplacer la simulation sous la loi conditionnelle par une étape de Hastings-Metropolis dont la loi cible est cette loi conditionnelle.

Vers les procédures MCMC adaptatives

Dans la mise en oeuvre de ces algorithmes, différents “degrés de liberté” laissés à l'utilisateur : par exemple,

- Hastings-Metropolis : quelle loi de proposition choisir ?
- Gibbs : quelle partition de l'espace d'état X ? quel ordre de mise à jour des composantes et avec quelle fréquence ?

⇒ définition d'algorithmes **adaptatifs** construits pour

- que l'algorithme apprenne de son propre comportement passé, comment modifier les **paramètres d'implémentation**
- et les modifie au fur et à mesure de son déroulement : chaque itération comporte une étape de mise à jour de la chaîne, et de mise à jour des paramètres d'implémentation.

Paramètres d'implémentation du SRW-HM (1/3) (Hastings-Metropolis à marche aléatoire symétrique)

Lorsque le noyau de proposition est de la forme

$$q(x,y) = q(y,x) \quad (\text{symétrie})$$

$$q(x,y) = q(y-x) \quad (\text{marche aléatoire})$$

une itération $X_n \rightarrow X_{n+1}$ devient :

- tirer $Z \sim q(z)dz$

- Poser

$$\begin{cases} X_{n+1} = X_n + Z & \text{avec probabilité } 1 \wedge \frac{\pi(X_n+Z)}{\pi(X_n)} \\ X_{n+1} = X_n & \text{sinon} \end{cases}$$

Paramètres d'implémentation du SRW-HM (1/3) (Hastings-Metropolis à marche aléatoire symétrique)

Lorsque le noyau de proposition est de la forme

$$q(x,y) = q(y,x) \quad (\text{symétrie})$$

$$q(x,y) = q(y - x) \quad (\text{marche aléatoire})$$

une itération $X_n \rightarrow X_{n+1}$ devient :

- tirer $Z \sim q(z)dz$

- Poser

$$\begin{cases} X_{n+1} = X_n + Z & \text{avec probabilité } 1 \wedge \frac{\pi(X_n+Z)}{\pi(X_n)} \\ X_{n+1} = X_n & \text{sinon} \end{cases}$$

En particulier, toute configuration $X_n + Z$ qui augmente la valeur de la densité cible est systématiquement acceptée.

Paramètres d'implémentation du SRW-HM (1/3) (Hastings-Metropolis à marche aléatoire symétrique)

Lorsque le noyau de proposition est de la forme

$$q(x,y) = q(y,x) \quad (\text{symétrie})$$

$$q(x,y) = q(y-x) \quad (\text{marche aléatoire})$$

une itération $X_n \rightarrow X_{n+1}$ devient :

- tirer $Z \sim q(z)dz$

- Poser

$$\begin{cases} X_{n+1} = X_n + Z & \text{avec probabilité } 1 \wedge \frac{\pi(X_n+Z)}{\pi(X_n)} \\ X_{n+1} = X_n & \text{sinon} \end{cases}$$

En particulier, toute configuration $X_n + Z$ qui augmente la valeur de la densité cible est systématiquement acceptée.

↔ En pratique, quel choix pour la loi de proposition q ?

Paramètres d'implémentation du SRW-HM (2/3)

Quelques réponses dans le cas d'une loi de proposition $\mathcal{N}_d(0, \Sigma)$:

- choisir pour matrice de var-covar: $\Sigma \propto \Sigma_\pi$.
- avec un facteur d'échelle qui tient compte de la dimension

$$\Sigma = \frac{2.38^2}{d} \Sigma_\pi$$

Paramètres d'implémentation du SRW-HM (2/3)

Quelques réponses dans le cas d'une loi de proposition $\mathcal{N}_d(0, \Sigma)$:

- choisir pour matrice de var-covar: $\Sigma \propto \Sigma_\pi$.
- avec un facteur d'échelle qui tient compte de la dimension

$$\Sigma = \frac{2.38^2}{d} \Sigma_\pi$$

Réponses apportées

- sous des conditions plus ou moins restrictives sur la densité cible.
- pour différents algorithmes de type Hastings-Metropolis.
- basées sur la vitesse de convergence optimale de "limites diffusives" de la chaîne (renormalisée).

Paramètres d'implémentation du SRW-HM (3/3)

En pratique, la matrice de var-covar Σ_π est inconnue \implies coupler simulation et estimation

Notons

- θ_n l'estimation courante de la matrice Σ_π .
- P_θ le noyau de transition d'une chaîne SRW-HM de loi de proposition $\mathcal{N}_d(0, \theta)$

Une itération $(\theta_n, X_n) \rightarrow (\theta_{n+1}, X_{n+1})$ devient

- **Simuler** $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$
- **Mettre à jour l'estimation** de Σ_π par la formule

$$\theta_{n+1} = \theta_n + \frac{1}{n+1} \left((X_{n+1} - \mu_{n+1})(X_{n+1} - \mu_{n+1})^T - \theta_n \right)$$

Algorithme de Wang-Landau (1/5)

Algorithme proposé *en dynamique moléculaire* pour être robuste à la multimodalité de la loi cible

- Choisir une partition $\mathcal{X}_1, \dots, \mathcal{X}_d$ du support \mathcal{X} de π .
- Au lieu d'échantillonner π , échantillonnons π_\star une loi qui va donner le même poids à tous les sous-espaces \mathcal{X}_i .

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

où $\theta_\star = (\theta_\star(1), \dots, \theta_\star(d))$ est un **vecteur de poids** i.e. $\sum_i \theta_\star(i) = 1$ and $\theta_\star(i) \geq 0$

$$\theta_\star(i) = \int_{\mathcal{X}_i} \pi(\mathrm{d}x)$$

Algorithme de Wang-Landau (1/5)

Algorithme proposé *en dynamique moléculaire* pour être robuste à la multimodalité de la loi cible

- Choisir une partition $\mathcal{X}_1, \dots, \mathcal{X}_d$ du support \mathcal{X} de π .
- Au lieu d'échantillonner π , échantillonnons π_\star une loi qui va donner le même poids à tous les sous-espaces \mathcal{X}_i .

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

où $\theta_\star = (\theta_\star(1), \dots, \theta_\star(d))$ est un **vecteur de poids** i.e. $\sum_i \theta_\star(i) = 1$ and $\theta_\star(i) \geq 0$

$$\theta_\star(i) = \int_{\mathcal{X}_i} \pi(\mathrm{d}x)$$

↪ **MAIS:**

- Les poids $\theta_\star(i)$ sont inconnus en pratique : il faut les apprendre !
- La loi cible π_\star n'est pas π donc il faudra corriger les tirages (mécanisme de repondération) pour obtenir une approximation de π .

Algorithme de Wang-Landau (2/5)

Wang-Landau est un algorithme de Monte Carlo qui **simultanément**

- échantillonne des points $\{X_k, k \geq 0\}$: $X_{k+1} \sim P_{\theta_k}(X_k, \cdot)$ où P_θ est un noyau de transition admettant π_θ pour loi invariante

$$\pi_\theta(dx) \propto \sum_{i=1}^d \frac{\pi(dx)}{\theta(i)} \mathbb{1}_{\mathcal{X}_i}(x).$$

- apprend le vecteur de poids θ_* : si $X_{k+1} \in \mathcal{X}_i$, une mise à jour possible :

$$\begin{cases} \theta_{k+1}(i) = \theta_k(i) + \gamma_{k+1} \theta_k(i)(1 - \theta_k(i)) \\ \theta_{k+1}(j) = \theta_k(j) - \gamma_{k+1} \theta_k(j)\theta_k(i) & j \neq i \end{cases}$$

Algorithme de Wang-Landau (2/5)

Wang-Landau est un algorithme de Monte Carlo qui **simultanément**

- échantillonne des points $\{X_k, k \geq 0\}$: $X_{k+1} \sim P_{\theta_k}(X_k, \cdot)$ où P_{θ} est un noyau de transition admettant π_{θ} pour loi invariante

$$\pi_{\theta}(dx) \propto \sum_{i=1}^d \frac{\pi(dx)}{\theta(i)} \mathbb{1}_{\mathcal{X}_i}(x).$$

- apprend le vecteur de poids θ_* : si $X_{k+1} \in \mathcal{X}_i$, une mise à jour possible:

$$\begin{cases} \theta_{k+1}(i) = \theta_k(i) + \gamma_{k+1} \theta_k(i)(1 - \theta_k(i)) \\ \theta_{k+1}(j) = \theta_k(j) - \gamma_{k+1} \theta_k(j)\theta_k(i) & j \neq i \end{cases}$$

- construit une approximation de π

grosso modo:
$$\frac{1}{n} \sum_{k=1}^n \delta_{X_k} \approx \pi_{\theta_*} \implies \frac{1}{n} \sum_{k=1}^n \left(\sum_{i=1}^d \theta_*(i) \mathbb{1}_{X_k \in \mathcal{X}_i} \right) \delta_{X_k} \approx \pi$$

$$\frac{d}{n} \sum_{k=1}^n \left(\sum_{i=1}^d \theta_k(i) \mathbb{1}_{\mathcal{X}_i}(X_k) \right) f(X_k) \xrightarrow{\text{p.s.}} \int f(x) \pi(dx)$$

Algorithme de Wang-Landau (3/5)

Densité cible: $\pi(x_1, x_2) \propto \exp(-\beta \mathcal{H}(x_1, x_2)) \mathbb{1}_{[-R, R]}(x_1)$

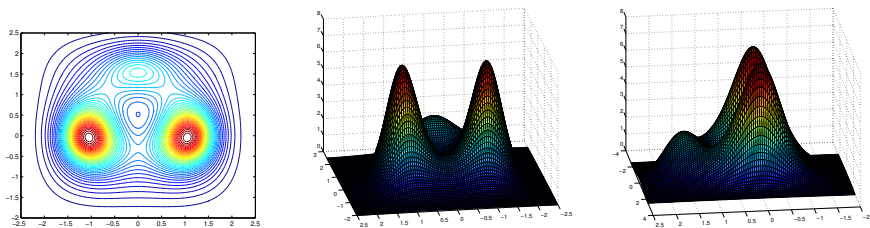
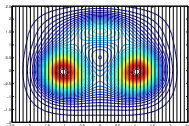


FIG.: [left] Courbes de niveau du potentiel \mathcal{H} . [center, right] Densité π à une constante près

Plus β est grand, plus grand est le ratio entre le poids de la strate contenant le mode et celui de la strate contenant la vallée

Algorithme de Wang-Landau (4/5)



$R = 2.4$. $d = 48$ strata, partition selon l'axe des x .

P_θ est un noyau de Hastings-Metropolis de loi de proposition $\mathcal{N}(0, (2R/d)^2 I)$ et loi cible π_θ . Chaîne initialisée en $X_0 = (-1, 0)$.

Suite de pas: $\gamma_t \sim c/t^{0.8}$.

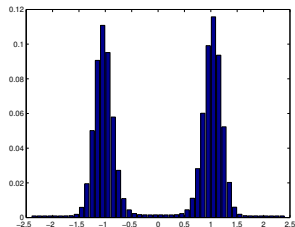
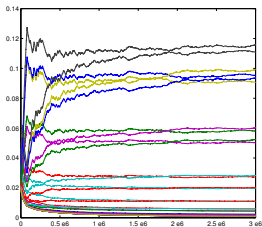


FIG.: [left] Estimation des poids $(\theta_t(i))_t$. [right] Valeur limite $\theta_\star(i)$

Algorithme de Wang-Landau (5/5)

Evolution de la première composante x_1 de $(X_t)_t$, quand X_t est la chaîne de Wang-Landau (gauche) et la chaîne de Hastings-Metropolis (droite).

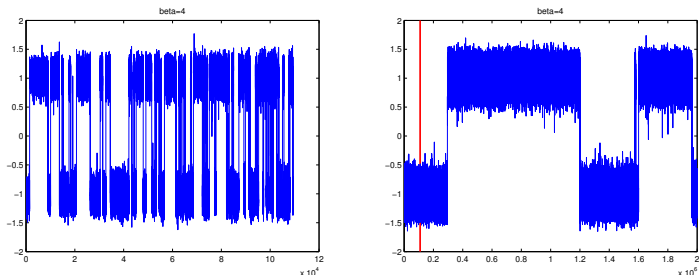


FIG.: [left] Wang Landau, $T = 110\,000$. [right] Hastings Metropolis, $T = 2 \times 10^6$; trait rouge à $x = 110\,000$

Conclusion

- ④ Les algorithmes MCMC classiques (non adaptatifs) :

$$\mathbb{E}[\phi(X_{n+1})|\mathcal{F}_n] = \int P(X_n, dy)\phi(y)$$

avec $\pi P = \pi$.

Conclusion

- ① Les algorithmes MCMC classiques (non adaptatifs) :

$$\mathbb{E}[\phi(X_{n+1})|\mathcal{F}_n] = \int P(X_n, dy)\phi(y)$$

avec $\pi P = \pi$.

- ② Les algorithmes MCMC adaptatifs

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

avec

- un processus d'adaptation θ_n tel que $\theta_{n+1} = \Xi_n(\theta_n, X_{n+1}, U)$
- une famille de noyaux de transition $\{P_\theta, \theta \in \Theta\}$
- avec $\pi_\theta P_\theta = \pi_\theta$ pour tout θ mais PAS nécessairement $\pi_\theta = \pi$.

↔ A quelles conditions le processus adapté $\{X_n, n \geq 0\}$ converge-t-il vers la loi cible π ?

II. Convergence des algorithmes MCMC adaptatifs

Résultats obtenus en collaboration avec E. MOULINES (TELECOM ParisTech)
et P. PRIOURET (Paris 6).

Par “convergence”, on entend

- convergence des marginales

$$\mathbb{E}[\phi(X_n)] \rightarrow \int \phi d\pi \quad \text{pour une famille de fonctions } \phi$$

- existence d'une loi des grands nombres pour une famille de fonctions (bornées ou pas)
- existence d'un théorème central limite.

Nous apportons des réponses dans le cas

- $\pi P_\theta = \pi$ pour tout θ
- $\pi_\theta P_\theta = \pi_\theta$ et $\pi_\theta \neq \pi$

sous forme de conditions **suffisantes** sur

- le noyau de transition P_θ
- et la façon dont est faite l'adaptation.

L'adaptation peut détruire la convergence

- Considérons la famille de noyaux de transition sur $\{0,1\}$:

$$P_\theta = \begin{pmatrix} 1 - \theta & \theta \\ \theta & 1 - \theta \end{pmatrix} \quad \theta \in (0,1)$$

Alors pour tout $\theta \in (0,1)$, $\pi P_\theta = \pi$ avec $\pi = (1/2; 1/2)$.

- Fixons $t_0, t_1 \in (0,1)$. On définit le processus adaptatif $\{X_n, n \geq 0\}$ par

$$\begin{cases} X_{n+1} \sim P_{\theta_n}(X_n, \cdot) \\ \theta_{n+1} = t_{X_{n+1}} \end{cases}$$

Alors le noyau de transition de ce processus adaptatif est

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

qui a pour mesure invariante $\pi \propto (t_1, t_0)$.

Conditions pour la convergence des marginales

On écrit

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

Conditions pour la convergence des marginales

On écrit

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

► [1] condition d'ergodicité des noyaux de transition

“Habituellement”, les noyaux de transition $\{P_\theta, \theta \in \Theta\}$ sont géométriquement ergodiques :

$$\sup_{f, \|f\|_\infty \leq 1} |P_\theta^n f(x) - \pi(f)| \leq C_\theta \rho_\theta^n V(x) \quad \rho_\theta \in (0,1)$$

MAIS la vitesse de convergence peut dépendre de θ de telle sorte que

$$\rho_\theta \rightarrow 1 \quad \text{quand } \theta \rightarrow \partial\Theta .$$

Par suite, il est nécessaire de contrôler la vitesse à laquelle $\theta_n \rightarrow \partial\Theta$

Conditions pour la convergence des marginales

On écrit

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

► [2] condition sur le mécanisme d'adaptation puisque

$$\left| \mathbb{E}\left[f(X_n) - P_{\theta_{n-N}}^N f(X_{n-N})\right] \right| \leq \sum_{j=1}^{N-1} (N-j) \mathbb{E} \left[\underbrace{\sup_x \left\| P_{\theta_{n-N+j}}(x, \cdot) - P_{\theta_{n-N+j-1}}(x, \cdot) \right\|_{\text{TV}}}_{\text{"distance" entre deux noyaux successifs}} \right]$$

Par la suite, l'adaptation doit être "décroissante".

Adaptation et Ergodicité

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

► Exemple: $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$ $\theta_n = n^{-1/4}$ $P_\theta = \begin{pmatrix} 1 - \theta & \theta \\ \theta & 1 - \theta \end{pmatrix}$

Dans ce cas, puisque $\theta_n \rightarrow 0$

$$\left| \mathbb{E}\left[P_{\theta_{n-N}}^N f(X_{n-N}) - \pi(f)\right] \right| (\leq) = |1 - 2\theta_{n-N}|^N \rightarrow 1 \quad \text{pour } N \text{ fixé}$$

Adaptation et Ergodicité

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

► Exemple: $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$ $\theta_n = n^{-1/4}$ $P_{\theta} = \begin{pmatrix} 1 - \theta & \theta \\ \theta & 1 - \theta \end{pmatrix}$

Donc prendre N fonction de n

$$\left| \mathbb{E} \left[P_{\theta_{n-N_n}}^{N_n} f(X_{n-N_n}) - \pi(f) \right] \right| \leq |1 - 2\theta_{n-N_n}|^{N_n} \rightarrow 0$$

et l'adaptation doit être telle que

$$\sum_{j=1}^{N_n-1} (N_n - j) \mathbb{E} \left[\underbrace{\sup_x \left\| P_{\theta_{n-N_n+j}}(x, \cdot) - P_{\theta_{n-N_n+j-1}}(x, \cdot) \right\|_{\text{TV}}}_{\text{"distance" entre deux noyaux consécutifs}} \right] \rightarrow 0$$

↔ La "vitesse" d'adaptation dépend du comportement ergodique des noyaux

Résultat

Si

- $\exists \pi$ tq $\pi P_\theta = \pi$ pour tout $\theta \in \Theta$.
- (**Ergodicité**) pour tout $\epsilon > 0$, il existe une suite positive croissante $\{r_\epsilon(n), n \geq 0\}$ tq $\limsup_{n \rightarrow \infty} r_\epsilon(n)/n = 0$ et

$$\limsup_{n \rightarrow \infty} \mathbb{E} \left[\left\| P_{\theta_{n-r_\epsilon(n)}}^{r_\epsilon(n)}(X_{n-r_\epsilon(n)}, \cdot) - \pi \right\|_{\text{TV}} \right] \leq \epsilon .$$

- (**Adaptation décroissante**) Pour tout $\epsilon > 0$,

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

Alors

$$\sup_{\{\phi, \|\phi\|_\infty \leq 1\}} |\mathbb{E}[\phi(X_n)] - \int \phi d\pi| \rightarrow 0 .$$

En pratique, comment vérifier les conditions?

Dans un contexte MCMC et plus précisément Hastings-Metropolis :

- 1 la condition d'ergodicité qui nécessite le contrôle de $\|P_\theta^N(x, \cdot) - \pi\|_{\text{TV}}$ s'obtient
 - en vérifiant une condition de **drift** : $P_\theta V \leq \lambda_\theta V + b_\theta \mathbb{1}_{\mathcal{C}_\theta}$
 - en vérifiant une condition de **minoration** : $P_\theta(x, \cdot) \geq \epsilon_\theta \mathbb{1}_{\mathcal{C}_\theta}(x) \nu_\theta(\cdot)$
 - et en utilisant des résultats sur **le calcul explicite de bornes d'ergodicité** à partir des quantités $b_\theta, \lambda_\theta, \epsilon_\theta, \mathcal{C}_\theta$.

En pratique, comment vérifier les conditions?

Dans un contexte MCMC et plus précisément Hastings-Metropolis :

- 1 la condition d'ergodicité qui nécessite le contrôle de $\|P_\theta^N(x, \cdot) - \pi\|_{\text{TV}}$ s'obtient
 - en vérifiant une condition de **drift** : $P_\theta V \leq \lambda_\theta V + b_\theta \mathbb{1}_{\mathcal{C}_\theta}$
 - en vérifiant une condition de **minoration** : $P_\theta(x, \cdot) \geq \epsilon_\theta \mathbb{1}_{\mathcal{C}_\theta}(x) \nu_\theta(\cdot)$
 - et en utilisant des résultats sur **le calcul explicite de bornes d'ergodicité** à partir des quantités $b_\theta, \lambda_\theta, \epsilon_\theta, \mathcal{C}_\theta$.
- 2 la condition d'adaptation décroissante s'obtient en établissant une inégalité de la forme

$$\|P_\theta(x, \cdot) - P_{\theta'}(x, \cdot)\|_{\text{TV}} \leq C d(\theta, \theta') .$$

Doit-on avoir convergence du processus d'adaptation ?

Selon la situation

- 1 **Non** : dans la situation que l'on vient d'étudier, la seule condition qui apparaît porte sur
 - $d(\theta_n, \theta_{n-1}) \rightarrow 0$
 - le contrôle de la vitesse à laquelle θ_n s'approche du bord de l'ensemble $\partial\Theta$.

Mais c'est une situation particulière dans laquelle les noyaux P_θ ont TOUS même mesure invariante.

Doit-on avoir convergence du processus d'adaptation ?

Selon la situation

- ① **Non** : dans la situation que l'on vient d'étudier, la seule condition qui apparaît porte sur
- $d(\theta_n, \theta_{n-1}) \rightarrow 0$
 - le contrôle de la vitesse à laquelle θ_n s'approche du bord de l'ensemble $\partial\Theta$.

Mais c'est une situation particulière dans laquelle les noyaux P_θ ont TOUS même mesure invariante.

- ② **Oui** : dans la situation où chaque noyau P_θ a sa propre mesure invariante π_θ . Dans ce cas, il faut des conditions pour que

$$\pi_{\theta_n}(\phi) \xrightarrow{\text{p.s.}} \pi(\phi) \quad \text{par ex. pour l'obtention d'une LGN}$$

et ces conditions implicitement demandent la convergence de θ_n vers θ_* tel que $\pi_{\theta_*} = \pi$. En pratique, on n'a pas d'expression explicite de π_θ et cela s'établit à l'aide de conditions sur P_θ .

Conclusion

- ▶ Les procédures MCMC adaptatives sont basées sur la construction conjointe d'un processus d'adaptation et du processus d'intérêt.
- ▶ Les procédures d'adaptation peuvent détruire la convergence de l'algorithme vers la loi cible.

Conclusion

- ▶ Les procédures MCMC adaptatives sont basées sur la construction conjointe d'un processus d'adaptation et du processus d'intérêt.
- ▶ Les procédures d'adaptation peuvent détruire la convergence de l'algorithme vers la loi cible.
- ▶ On peut néanmoins montrer la convergence de procédures au sens de adaptatives
 - convergence des marginales
 - loi des grands nombres
 - théorème central limite

dès lors que celles-ci vérifient des conditions

- (i) d'ergodicité de chaque noyau de transition
- (ii) d'adaptation "décroissante"
- (iii) et dans certains cas - de convergence des noyaux P_{θ_n} vers un noyau P_{θ_*} .

Références

Méthodes MCMC adaptatives : méthodologie

- C. Andrieu, J. Thoms. *Statistics and Computing*, 2008.
- J.S. Rosenthal. *MCMC handbook*, 2009.
- Y. Atchadé, G. Fort, E. Moulines, P. Priouret. *Times Series book*, 2010.

Méthodes MCMC adaptatives : résultats de convergence

- G.O. Roberts, J.S. Rosenthal. *J. Appl. Probab.* 2007.
- Y. Atchadé, G. Fort. *Bernoulli*, 2010.
- G. Fort, E. Moulines, P. Priouret. *Convergence of adaptive MCMC algorithms: ergodicity and law of large numbers*. *Ann. Stat.* 2012
- G. Fort, E. Moulines, P. Priouret and P. Vandekerkhove. A Central Limit Theorem for Adaptive and Interacting Markov Chain. *Bernoulli*, 2013.

Analyse de convergence d'algorithmes adaptatifs et en interaction

Adaptive Metropolis

- C. Andrieu, E. Moulines. *Ann. Appl. Probab.*, 2006.
- E. Saksman, M. Vihola, *On the ergodicity of the adaptive Metropolis algorithm on unbounded domains*. *Ann. Appl. Probab.* 2010

Wang-Landau

- G. Fort, B. Jourdain, E. Kuhn, T. Lelièvre and G. Stoltz. *Convergence of the Wang-Landau algorithm*. *Accepted for publication in Mathematics of Computation*, 2014. arXiv math.PR 1207.6880.
- G. Fort, B. Jourdain, E. Kuhn, T. Lelièvre and G. Stoltz. *Efficiency of the Wang-Landau algorithm* *Accepted for publication in Applied Mathematics Research Express*, 2014. arXiv 1310.6550

Analyse de convergence d'algorithmes adaptatifs et en interaction (suite)

Equi-Energy sampler

- C. Andrieu, A. Jasra, A. Doucet, P. DelMoral. *On non-linear Markov chain Monte Carlo*, Bernoulli 2011.
- G. Fort, E. Moulines, P. Priouret. *Convergence of adaptive MCMC algorithms: ergodicity and law of large numbers*. Ann. Stat. 2012
- G. Fort, E. Moulines, P. Priouret and P. Vandekerkhove. *A Central Limit Theorem for Adaptive and Interacting Markov Chain*. Bernoulli, 2013.
- A. Schreck, G. Fort and E. Moulines. *Adaptive Equi-energy sampler : convergence and illustration*. TOMACS, 2013.