Des Résultats théoriques sur les algorithmes Monte Carlo par chaînes de Markov

Jeffrey S. Rosenthal

Department of Statistics, University of Toronto 100 St. George Street, room 6018, Toronto, Ontario, Canada M5S 3G3.

Résumé :

Nous discutons des algorithmes Monte Carlo par chaînes de Markov (MCMC), et surtout de comment les résultats théoriques peuvent aider avec leur utilisation. Nous présentons un théorème qui donne des bornes quantitatives sur leur distance à la distribution stationnaire. Nous discutons aussi des algorithmes adaptatifs, et d'un théorème qui donne des conditions qui assurent qu'ils convergent vers la distribution stationnaire (ce qui n'est pas vrai en général).

Résumé en anglais :

We describe Markov chain Monte Carlo (MCMC) algorithms, with an emphasis on ways in which theoretical results can help with their implementation. We present a theorem which gives quantitative bounds on their distance to stationarity. We also discuss adaptive MCMC algorithms, and a theorem which gives conditions guaranteeing their convergence to stationarity (which does not hold in general).

Mots clés : Algorithmes Monte Carlo, chaînes de Markov, taux de convergence, algorithmes adaptatifs.

Les algorithmes Monte Carlo par chaînes de Markov (MCMC) sont treès bien connus en physique statistique (Metropolis et al., 1953), en analyse des images (Geman et Geman, 1984), et en inférence statistique (Hastings, 1970 ; Gelfand et Smith, 1990 ; Tierney, 1994). On s'intéresse aux échantillons d'une distribution $\pi(\cdot)$ compliquée sur une espace $\mathcal{X}$ en grande dimension. On définit un noyau de transition $P(x, \cdot)$ qui laisse invariant $\pi(\cdot)$, on commence avec une valeur initiale X_0, on crée $X_n \sim P(X_{n-1}, \cdot)$ pour $n = 1, 2, 3, \ldots$, et on « espère » que pour grand n, la loi de X_n est presque $\pi(\cdot)$.

Mais, ça pose plusieurs questions théoriques, surtout (a) comment peut-on savoir si la loi de X_n est presque $\pi(\cdot)$, et (b) est-ce qu'on peut redéfinir $P(x, \cdot)$ pour avoir une plus vite convergence.

Pour la question (a), nous discuterons du théorème suivant (Rosenthal, 1995, 2002 ; Douc et al., 2004 ; Jones et Hobert, 2001). Définissons

$$\|\mathcal{L}(X_n) - \pi(\cdot)\| \;=\; \sup_{A \subseteq \mathcal{X}} |\mathbf{P}(X_n \in A) - \pi(A)| \,.$$

Supposons qu'il existe une fonction $V : \mathcal{X} \to [0, \infty)$, et $\lambda < 1$ et $\Lambda < \infty$, qui satisfaisent la *condition de drift* :

$$\mathbf{E}\Big(V(X_n) \,|\, X_{n-1} = x\Big) \;\leq\; \lambda\, V(x) + \Lambda\,, \qquad x \in \mathcal{X}\,. \tag{1}$$

Supposons aussi qu'il existe $D > 0$, $\epsilon > 0$, et n'importe quelle distribution $\nu(\cdot)$ sur $\mathcal{X}$, qui satisfaisent la *condition de minorisation* (où de *small set*) :

$$P(x, \cdot) \;\geq\; \epsilon\, \nu(\cdot)\,, \qquad \forall x \text{ avec } V(x) \leq D\,. \tag{2}$$

Théorème 1 *Si les conditions (1) et (2) sont satisfaites, avec* $D > \frac{2\Lambda}{1-\lambda}$*, alors pour n'importe quel entier* $0 \leq j \leq n$*,*

$$\|\mathcal{L}(X_n) - \pi(\cdot)\| \;\leq\; (1-\epsilon)^j \;+\; \alpha^{-n+j-1}\Delta^j \left(1 + \frac{\Lambda}{1-\lambda} + \mathbf{E}(V(X_0))\right), \tag{3}$$

où $\alpha = \frac{1+D}{1+2\Lambda+\lambda D} > 1$ *et* $\Delta = 1 + 2(\lambda D + \Lambda)$*.*

Théorème 1 nous donne la possibilité de vérifier que, par exemple, $\|\mathcal{L}(X_n) - \pi(\cdot)\| < 0.01$ pour certains grands n. Pour un exemple spécifique et assez complexe en dimension 20, avec des données des joueurs de baseball, Théorème 1 à été utilisé (Rosenthal, 1996) pour prouver que $\|\mathcal{L}(X_n) - \pi(\cdot)\| < 0.01$ si $n = 140$, qui est une borne raisonnable et pratique.

Pour la question (b), plusieurs auteurs (Haario et al., 2001 ; Atchadé et Rosenthal, 2005 ; Andrieu et Moulines, 2006 ; Roberts et Rosenthal, 2007, 2006 ; Atchadé et Fort, 2008 ; Bai et al., 2008) ont considéré la possibilité des algorithmes *adaptatifs*. On imagine qu'il y a plusieurs noyaux $\{P_\gamma\}_{\gamma \in \mathcal{Y}}$, chacun desquels laisse stationnaire la distribution $\pi(\cdot)$. On veut choisir le « meilleur » $\gamma \in \mathcal{Y}$. On le fait du style aléatoire : pour l'itération n, on utilise le noyau P_{Γ_n} où $\Gamma_n \in \mathcal{Y}$ dépend de l'histoire $X_0, \ldots, X_{n-1}$ et $\Gamma_0, \ldots, \Gamma_{n-1}$. On espère le faire d'une manière qui cherche une bonne valeur de γ, en considérant d'autres critères (comme le taux d'acceptation des propositions Metropolis, etc.).

Ces algorithmes adaptatifs ne sont plus Markoviens, et en général ils ne vont pas converger vers $\pi(\cdot)$ (voir par exemple la présentation interactive de Rosenthal, 2004). Mais, du point de vue positif, on a le théorème suivant (Roberts et Rosenthal, 2006). Définissons

$$M_\epsilon(x, \gamma) \;=\; \inf\{n \geq 1 : \|P_\gamma^n(x, \cdot) - \pi(\cdot)\| \leq \epsilon\}\,.$$

Théorème 2 *Pour l'algorithme adaptatif, on a* $\lim_{n\to\infty} \|\mathcal{L}(X_n) - \pi(\cdot)\| = 0$*, si (i)* $\pi(\cdot)$ *est stationnaire pour chaque* P_γ *individuel, et (ii)* $\lim_{n\to\infty} \sup_{x\in\mathcal{X}} \|P_{\Gamma_{n+1}}(x, \cdot) - P_{\Gamma_n}(x, \cdot)\| = 0$ *en probabilité (« Diminishing Adaptation »), et (iii) pour chaque* $\epsilon > 0$*, les valeurs* $\{M_\epsilon(X_n, \Gamma_n)\}_{n=0}^\infty$ *sont bornées en probabilité (« Containment »).*

Pour des exemples complexes en grande dimension, on peut (Roberts and Rosenthal, 2006 ; Bai et al., 2008) parfois vérifier ces conditions et utiliser des algorithmes adaptatifs pour bien améliorer le taux de convergence vers $\pi(\cdot)$.

Bibliographie

C. Andrieu et E. Moulines (2006), On the ergodicity properties of some adaptive Markov Chain Monte Carlo algorithms. Ann. Appl. Prob. **16**, 1462–1505.

Y. Atchadé et G. Fort (2008), Limit theorems for some adaptive MCMC algorithms with subgeometric kernels. Prépublication.

Y.F. Atchadé et J.S. Rosenthal (2005), On Adaptive Markov Chain Monte Carlo Algorithms. Bernoulli **11**, 815–828.

Y. Bai, G.O. Roberts, et J.S. Rosenthal (2008), On the Containment Condition for Adaptive Markov Chain Monte Carlo Algorithms. Prépublication.

R. Douc, E. Moulines, et J.S. Rosenthal (2002), Quantitative bounds on convergence of time-inhomogeneous Markov Chains. Ann. Appl. Prob. **14**, (2004), 1643–1665.

A.E. Gelfand et A.F.M. Smith (1990), Sampling based approaches to calculating marginal densities. J. Amer. Stat. Assoc. **85**, 398–409.

S. Geman et D. Geman (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. on pattern analysis and machine intelligence **6**, 721–741.

H. Haario, E. Saksman, et J. Tamminen (2001), An adaptive Metropolis algorithm. Bernoulli **7**, 223–242.

W.K. Hastings (1970), Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97–109.

G.L. Jones et J.P. Hobert (2001), Honest exploration of intractable probability distributions via Markov chain Monte Carlo. Statistical Science **16**, 312–334.

N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, et E. Teller (1953), Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1091.

G.O. Roberts et J.S. Rosenthal (2007), Coupling and Ergodicity of Adaptive MCMC. J. Appl. Prob. **44**, 458–475.

G.O. Roberts et J.S. Rosenthal (2006), Examples of Adaptive MCMC. J. Comp. Graph. Stat., à paraître.

J.S. Rosenthal (1995), Minorization conditions and convergence rates for Markov chain Monte Carlo. J. Amer. Stat. Assoc. **90**, 558–566.

J.S. Rosenthal (1996), Convergence of Gibbs sampler for a model related to James-Stein estimators. Stat. and Comput. **6**, 269–275.

J.S. Rosenthal (2002), Quantitative convergence rates of Markov chains : A simple account. Elec. Comm. Prob. **7**, No. 13, 123–128.

J.S. Rosenthal (2004), Adaptive MCMC Java Applet. Disponible à :
http ://probability.ca/jeff/java/adapt.html

L. Tierney (1994), Markov chains for exploring posterior distributions (avec discussion). Ann. Stat. **22**, 1701–1762.