Les marches aléatoires et les algorithmes MCMC

Jeffrey S. Rosenthal

University of Toronto

jeff@math.toronto.edu

http ://probability.ca/jeff/

(CRM, Montréal, Jan 12, 2007)

Un processus stochastique

Qu'est-ce que c'est ?

- Une collection des instructions probabilistiques pour « quoi faire la prochaine fois ».
- Les instructions sont suivues en répétition.
- Après plusieurs répétitions, même des instructions simples peuvent produire des resultats très interessants.
- Plusieurs applications aux jeux, algorithmes aléatoires, et beaucoup plus.

Premier exemple : marche aléatoire simple

Tu paris \$1, en répétition. Chaque fois, tu gagnes \$1 avec probabilité p, ou perds \$1 avec probabilité $1 - p$. $(0 < p < 1)$

C'est-à-dire :

Tu commences avec une fortune initiale X_0.

Puis, pour $n = 1, 2, \ldots$, X_n est égale à $X_{n-1} + 1$ avec prob p, ou $X_{n-1} - 1$ avec prob $1 - p$.

Équivalence : $X_n = X_0 + Z_1 + Z_2 + \ldots + Z_n$, où les $\{Z_i\}$ sont indépendants, avec $\mathbf{P}[Z_i = +1] = p = 1 - \mathbf{P}[Z_i = -1]$.

[APPLET]

Marche aléatoire simple (continué)

Même cette exemple simple est très interessante :

- Distribution : $\frac{1}{2}(X_n - X_0 + n) \sim \text{Binomial}(n, p)$
- Distribution limitée : $\frac{1}{\sqrt{np(1-p)}}(X_n - X_0 - n(2p-1)) \approx \text{Normal}(0,1)$ (n grand) (TLC)
- Récurrence : $\mathbf{P}[\exists\ n \geq 1 : X_n = X_0] = 1$ ssi $p = 1/2$ (symétrique) (toujours vrai en dimension 2, mais pas en ≥ 3)
- Fluctuations : Si $p = 1/2$, le processus éventuelement touchera à n'importe quelle sequence $a_1, a_2, \ldots, a_\ell$.
- Martingale : Si $p = 1/2$, alors $\mathbf{E}(X_n \,|\, X_0, \ldots, X_{n-1}) = X_{n-1}$, c.à.d. le processus reste le même en moyenne.
  Si $p \neq 1/2$, alors $\{((1-p)/p)^{X_n}\}$ est martingale.

La ruine du jouer

Quelle est la probabilité que $X_n = 2X_0$ avant $X_n = 0$?

Exemple : $p = 0.492929$ (comme le jeux « craps »). [APPLET]

Impossible à resoudre avec des computations directes, parce que le nombre d'iterations n'est pas borné.

Mais, avec des analyses plus fort (p.e. des martingales), on trouve :

Jeux :	Symétrique	Craps	Roulette
$X_0 = 1$	50%	49.29%	47.37%
$X_0 = 10$	50%	42.98%	25.85%
$X_0 = 100$	50%	1 sur 18	1 sur 37,000
$X_0 = 500$	50%	1 sur 1.4 million	1 sur 10^{23}
$X_0 = 1,000$	50%	1 sur 10^{16}	1 sur 10^{48}

Évidence claire pour la loi des grands nombres !

La convergence en distribution

Exemple : marche aléatoire simple symétrique ($p = 1/2$), sauf forcé à rester dans $\mathcal{X} = \{0, 1, \ldots, 6\}$.

c.à.d. : si le processus essaye de quitter $\mathcal{X}$, alors l'étappe est rejetée, et le processus reste le même ($X_n = X_{n-1}$).

Qu'est-ce qui se passe après beaucoup d'iterations ? [APPLET]

La **distribution empirique (noir)** converge vers la distribution désirée (bleu).

Interessant ? Oui. Utile ? En effet !

Generalisation

Soit $\pi(\cdot)$ une distribution (cas discret) où densité (cas continue) sur une espace $\mathcal{X}$. [Avant : $\pi(\cdot) = \text{Uniform}\{1, 2, 3, 4, 5, 6\}$.]

De X_{n-1}, proposer Y_n (symétriquement). Accepter $(X_n = Y_n)$ avec probabilité $\min[1,\ \pi(Y_n)/\pi(X_{n-1})]$. Sinon, rejeter $(X_n = X_{n-1})$. ("Algorithme Metropolis", 1953) [APPLET]

Alors, "probablement", si B et M sont grand, alors $X_B \approx \pi(\cdot)$, et

$$\mathbf{E}_\pi(h) \ \approx \ \frac{1}{M} \sum_{i=B}^{B+M-1} h(X_i)\,.$$

"Markov Chain Monte Carlo" (MCMC). Très populaire en statistique, physique, science informatique, finance, et plus. La preuve ?

788,000 pages web en Google !

Comment évaluer les algorithmes MCMC ?

e.g. exemple de l'applet, mais avec $\pi\{2\} = 0.0001$. Proposer par

$$Y_n \sim \text{Uniform}\{x-\gamma, \ldots, x-1, x+1, \ldots, x+\gamma\},$$

$\gamma \in \mathbf{N}$. [Avant : $\gamma = 1$.]

Quels γ donnent la bonne convergence ? [APPLET]

$\gamma = 1$ (comme avant) : trop petit, alors ne bouge pas assez.

$\gamma = 50$: trop grand, alors trop de rejets.

$\gamma = 3$ ou 4 ou 5 : un compromis, qui marche très bien.

Facile, ici. Mais, comment dècider dans un exemple complex ...

Une application statistique typique

$\pi(\cdot)$ a la densité suivante sur $\mathbf{R}^{K+3}$:

$$f(\sigma_\theta^2, \sigma_e^2, \mu, \theta_1, \ldots, \theta_K) \;=$$
$$C\; e^{-b_1/\sigma_\theta^2} {\sigma_\theta^2}^{-a_1-1} e^{-b_2/\sigma_e^2} {\sigma_e^2}^{-a_2-1} e^{-(\mu-\mu_0)^2/2\sigma_0^2}$$
$$\times \prod_{i=1}^{K} \left[e^{-(\theta_i-\mu)^2/2\sigma_\theta^2}/\sigma_\theta\right] \times \prod_{i=1}^{K}\prod_{j=1}^{J} \left[e^{-(Y_{ij}-\theta_i)^2/2\sigma_e^2}/\sigma_e\right],$$

où K, J grand, $\{Y_{ij}\}$ données (connues), $a_1, a_2, b_1, b_2, \mu_0, \sigma_0^2$ paramétres (connues), et $C > 0$ est la constante de nomalisation.

[Posterieur pour le modèle Variance Components.]

Integration numerique : impossible (même pour calculer C).

Metropolis : Oui ! Proposer p.e. $Y_n \sim \text{Normal}(X_{n-1}, \sigma^2)$.

Mais, avec quel σ^2 ?

Approche théoretique #1 : couplage

Si nous pouvons construire, avec $\{X_n\}$, un autre processus $\{X'_n\}$ pour lequel $X'_n \sim \pi(\cdot)$ pour chaque n, alors :

$$\begin{aligned} |\mathbf{P}(X_n \in A) - \pi(A)| &= |\mathbf{P}(X_n \in A) - \mathbf{P}(X'_n \in A)| \\ &\leq \mathbf{P}(X_n \neq X'_n)\,. \end{aligned}$$

Si la construction a la proprieté que $\mathbf{P}(X_n \neq X'_n) \approx 1$ pour n grand, alors ça nous donne beaucoup d'information sur la convergence en distribution.

Possible, et il y a quelques succes avec des exemples compliqués. [R., JASA, 1995 ; Stat. Comput. 1996 ; Elec. Comm. Prob. 2002 ; JASA 2003 ; etc.]

Mais pas facile ! (Minorisations, drifts, ...) [article]

Approche théoretiques #2 : échelles optimales

Il existe des théorèmes qui dissent, dans certains contextes, quelle valeur de γ (ou σ^2) est optimale :

Si la distribution désirée a des composants i.i.d., alors pour l'algorithme Metropolis, c'est optimale d'avoir un taux d'acceptance de 0.234. [Roberts, Gelman et Gilks, Ann. Appl. Prob. 1994]

Pour l'algorithme Langevin, le taux optimal est 0.574. [Roberts et R., JRSSB 1998 ; Stat. Sci. 2001]

Parfois ces resultats generalisent, et parfois pas. [M. Bédard, 2006]

Mais, tous ces contexts sont trop specifiques pour des « vrais » exemples. Alors, quoi faire en practique ?

MCMC Adaptif

Idée : Demander a l'ordinateur de trouver des bons γ pour nous.

C.à.d., à chaque iteration n, l'ordinateur va choisir une valeur $\{\Gamma_n\}$ à utiliser pour γ. Essayer de faire « apprendre » à l'ordinateur quelles valeurs sont les meilleures. Pour l'applet, par exemple :

- Chaque fois que Y_n est accepté, alors $\Gamma_n = \Gamma_{n-1} + 1$ (alors γ augmente, et le taux d'acceptence diminue).
- Chaque fois que Y_n est rejeté, alors $\Gamma_n = \max(\Gamma_{n-1} - 1, 1)$ (alors γ diminue, et le taux d'acceptence augmente).

Logique, et naturale. Mais est-ce que ça marche ? [APPLET]

NON ! C'est un désastre !

Quand est-ce les algorithmes adaptifs convergent ?

[Roberts et R., 2004, 2005]

Théorème. Un algorithme adaptif converge si
(a) les taux de convergence sont tous bornées [condition téchnique ; il suffit que $\mathcal{X}$ est fini ou compacte] ; et
(b) l'adaptation diminue : $\mathbf{P}[\Gamma_n \neq \Gamma_{n-1}] \to 0$, ou plus generalement $\sup_{x \in \mathcal{X}} \|P_{\Gamma_{n+1}}(x, \cdot) - P_{\Gamma_n}(x, \cdot)\| \to 0$.

Alors, dans l'exemple de l'applet, si on ne change Γ_n qu'avec probabilité $p(n)$, et $p(n) \to 0$, alors ça va converger bien.

Autres exemples des algorithmes adaptifs

Autres exemples auxquels le théorème s'applique :

- Metropolis-Hastings avec $Y_n \sim MVN(X_{n-1},\, v_n(X_0, \ldots, X_{n-1}))$, pour des fonctiones appropriées v_n.
- Metropolis-within-Gibbs : chacune des 500 variables a sa propre variance σ_i^2 de sa propre Y_i, et l'ordinateur adapte chaque σ_i^2 sépérament. Et, ça marche !
- L'algorithme "Adaptive Metropolis" : $Y_n \sim MVN(X_{n-1},\, c\,\Sigma_n)$, où $c > 0$, et Σ_n est l'estime empirique de la covariance de $\pi(\cdot)$. Ça marche, même en dimension 200 (quand Σ_n a dimension vers 20,000).

Conclusion : Souvent, les algorithmes adaptifs marchent bien !

Résumé

Les processus aléatoires sont très interessants, et parfois très utiles.

- La répétition des instructions (simples ?) probabilistiques.
- Distributions, limites, récurrence, fluctuations, martingales, la ruine du jouer, ...
- MCMC (Metropolis etc.) pour converger en distribution.
- Approches théoretiques : couplage, échelles optimales.
- Algorithmes adaptifs : l'ordinateur choisi pour nous. Si on fait beaucoup d'attention, ça peut marcher bien.

Beaucoup de questions récherches interressantes !