Méthodes de Monte Carlo pour la Modélisation et le Calcul Intensif

Applications à la Physique Numérique et à la Biologie

Séminaire CIMENT GRID

Introduction aux Méthodes de Monte Carlo

Olivier François

Que proposent les méthodes de Monte Carlo ?

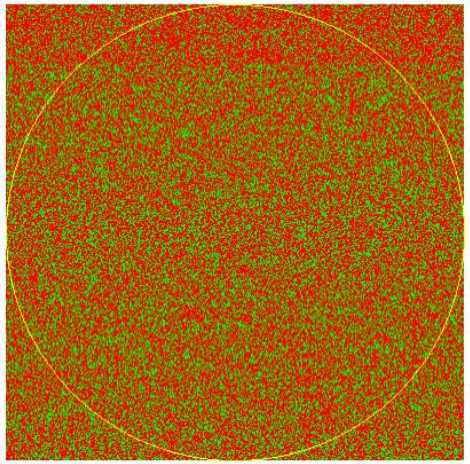
- Solutions approchées (et aléatoires) pour de nombreux problèmes
- Méthodes fondées sur la simulation informatique de variables aléatoires
- Applications en physique, en optimisation, en chimie, en imagerie, en statistique, en génétique, etc

Exemple : Calcul de volumes

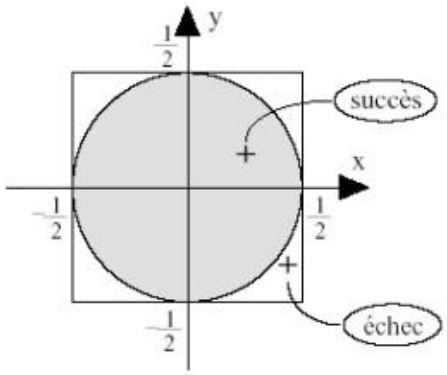
- Calcul d'intégrales en grandes dimensions

$$I = \int f(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

- Erreur en $O(1/\sqrt{n})$ contre $O(n^{-1/d})$ pour les méthodes fondées sur n points
- Calcul de π avec $n = 10000$ tirages $\Rightarrow \pi \approx 3.146$



$$\pi = 4\,\frac{\mathcal{A}(\text{cercle})}{\mathcal{A}(\text{carré})}$$



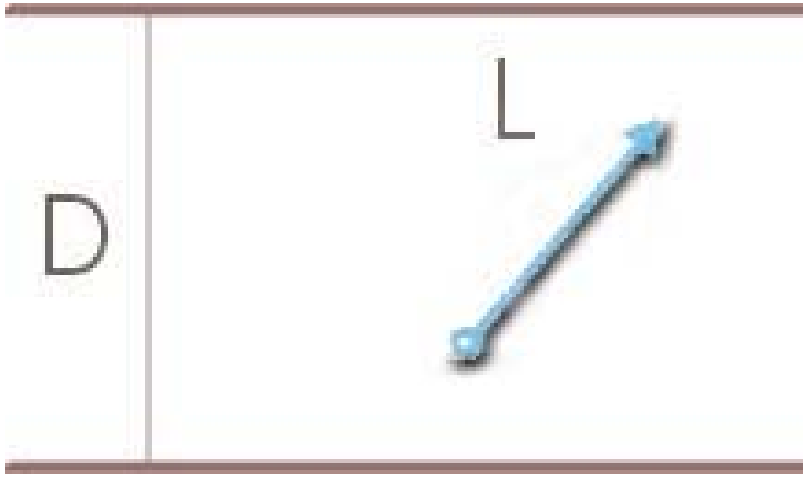
$$\pi = \lim_{n\to\infty} 4\,\frac{n_S}{n}$$

Avez vous vu Monte Carlo ?

- G. Leclerc (Buffon) écrit l'un des premiers traité de calcul des probabilités en liaison avec le calcul différentiel et intégral (1733). Expérience de l'aiguille

$$\mathrm{Proba}(\mathrm{L'aiguille\ coupe\ une\ ligne\ horizontale}) = \frac{2L}{\pi D}$$

- W. Gosset : tests statistiques (Student)
- Monte Carlo : Metropolis, Fermi et Ulam pour la simulation de la diffusion des neutrons dans un materiau fissible (Manhattan Project).



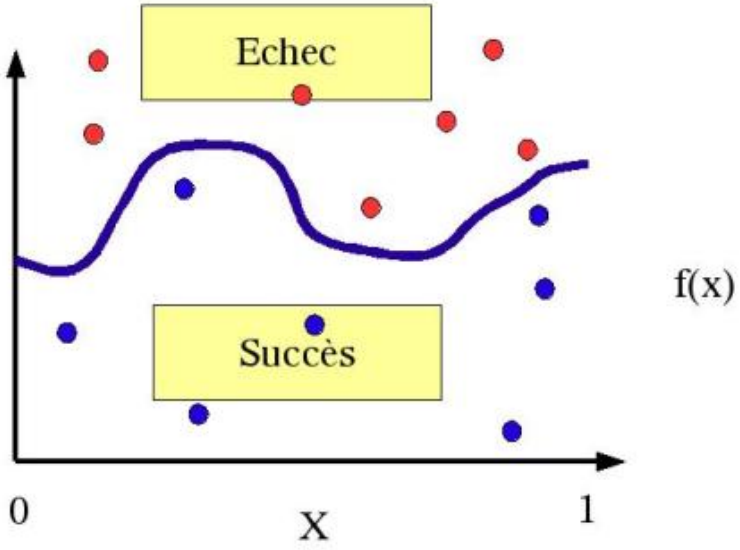
Principe

- Tirages de loi uniforme $U_1, \ldots, U_n$

$$I[f] \approx \frac{1}{n} \sum_i f(U_i)$$

- Erreur $= \sqrt{(I[f^2] - I^2)/n}$

- Tirages dans le rectangle $(0,1) \times (0,c) \rightarrow (U_1, cV_1) \ldots, (U_n, cV_n)$
- Erreur $= \sqrt{(I(c-I))/n}$



$$I \approx \frac{c}{n} \{\text{Nombre de } \bullet \text{ t.q. } cV_i < f(U_i)\}$$

Améliorations

- Tirages non uniformes : $X_1, \ldots, X_n$

$$I \approx \frac{1}{n} \sum_i \frac{f(X_i)}{q(X_i)}, \quad X_i \sim q$$

- Echantillonnage selon l'importance

$$I = \int \frac{f(x)}{q(x)} q(x) dx$$

Fonction de partition

- En mécanique statistique, l'espace des configurations d'un système peut être exponentiellement croissant (ex : aimants)

$$x \in \{-1, +1\}^S$$

  et une configuration est typiquement distribuée selon la loi de Boltzmann.

- on souhaite calculer par exemple la moyenne d'une variable observable φ

$$E[\varphi(X)] = \sum_x \varphi(x) \, \frac{\exp(-\beta H(x))}{Z_\beta}$$

- Mais

$$Z_\beta = \sum_x \exp(-\beta H(x))$$

  est en général incalculable.

- L'algorithme de Metropolis permet la simulation sans calculer Z_β

$$H(x) = \sum_{ij} x_i x_j + h \sum_i x_i$$

La dynamique de Metropolis : une méthode de simulation approchée

- Chaîne de Markov dont le régime stationnaire correspond à la loi souhaitée

$$\pi_\beta(x) = \frac{\exp(-\beta H(x))}{Z_\beta}$$

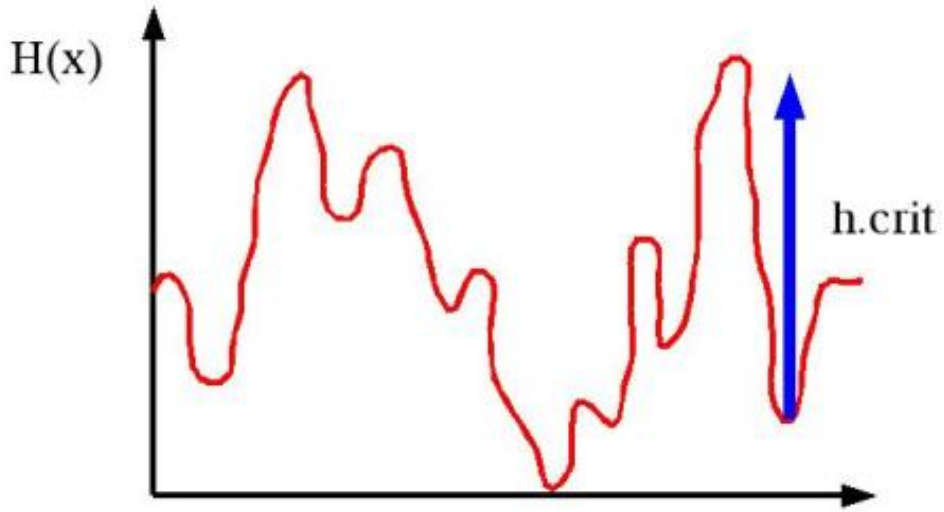
- dont les probabilités de transitions ne dépendent pas de Z_β

$$p(x,y) = \begin{cases} 1 & \text{si } \pi_\beta(y) \geq \pi_\beta(x) \\ \frac{\pi_\beta(y)}{\pi_\beta(x)} & \text{sinon.} \end{cases}$$

- Equations de bilan détaillé

$$\pi_\beta(x)p(x,y) = \pi_\beta(y)p(y,x)$$

- Qualité de l'approximation : INCONNUE en général



Temps d'atteinte de l'équilibre $\approx \exp(\beta h_c)$

Méthode de Monte Carlo en statistique

- Pour un jeu de données D, le calcul de la vraisemblance

$$L(\theta) = p(D/\theta)$$

  permet d'estimer le paramètre θ.

- En statistique bayésienne, on introduit souvent une variable latente H de sorte que $p(D/H)$ est calculable et $p(H/\theta)$ connu

- Chaînes de Markov cachées, filtrage (Kalman) non linéaire, reconstruction d'images, etc

- Difficulté : le calcul direct de la vraisemblance est infaisable

$$L(\theta) = \int_{\mathcal{H}} p(D/H)p(H/\theta)dH$$

Principe de la méthode de Monte Carlo

- Calcul approché de la vraisemblance : *Echantillonnage selon l'importance*

$$L(\theta) \propto \frac{1}{n} \sum_{t=1}^{n} \frac{p(H^{(i)}/\theta)}{p(H^{(i)}/\theta_0)}$$

  où θ_0 est une valeur de référence (estimation grossière).

- $p(H/\theta)$ doit être calculée de manière exacte

- Les variables latentes $H^{(i)}$ sont simulées selon la loi d'importance

$$q(H) = p(H/D;\theta_0) \propto p(D/H)p(H/\theta_0).$$

Simulation de la loi d'importance

- Méthode de *Métropolis-Hastings* : Markov chain Monte Carlo (McMC)
- On fait évoluer H de manière itérative en ne modifiant en général qu'une seule de ses composantes (transitions locales) conditionnellement aux autres : Echantillonnage de Gibbs.
- *Méthode de rejet* fondée sur le calcul de $p(D/H)$.
- Il est extrèmement difficile de contrôler la vitesse de convergence de cette méthode.

Un exemple en biologie de la conservation

Collaboration avec S. Manel, E. Bellemain (LECA), J. Swenson (NLH, Norvège)

- Espèce *Ursus arctos* (Ours brun de Scandinavie) : 516 individus séquencés, 19 marqueurs génétiques neutres de type *STR* (Short Tandem Repeats)
- Population décimée jusqu'en 1930, protégée depuis, et reconstituée à partir de 4 foyers maternels originels bien connus géographiquement.
- Objectif : Estimer l'évolution de la structure de la population et la dispersion spatiale des individus et des gènes de génération en génération.

Modélisation statistique

- Modèle aléatoire de généalogie décrivant les relations de parenté entre individus : le *coalescent* (Kingman, 1982).
- Événements structurants : mutations (μ), migrations (m) décrits par des paramètres scalaires

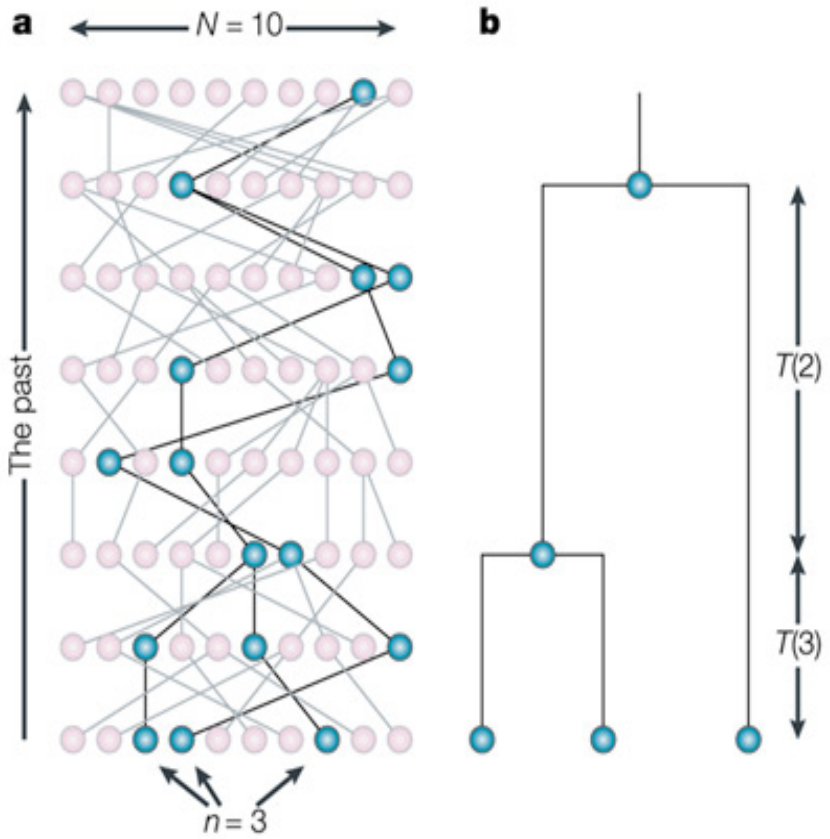
$$\theta = 4N\mu \qquad \text{et} \qquad \mathcal{M} = 4Nm$$

- La véritable généalogie n'est pas connue : Modèle à données manquantes

Arbre de coalescence

- Modèle de généalogie de Kingman

$$p(G/\theta) = \frac{1}{\theta^{n-1}} \exp(-\sum_{k=2}^{n} \lambda_k \frac{t_k}{\theta}), \quad \lambda_k = k(k-1)/2$$



Estimation par une méthode de maximum de vraisemblance

- Calcul de la vraisemblance

$$L(\theta) = p(D/\theta)$$

  D représente l'ensemble des observations moléculaires.

- Difficulté : le calcul direct de la vraisemblance est infaisable

$$L(\theta) = \int_{\mathcal{G}} p(D/G)p(G/\theta)dG$$

- Nécessité d'intégrer sur l'ensemble de toutes les généalogies $\mathcal{G}$.

Échantillon d'arbres tirés selon le modèle du coalescent

Lors d'un échantillonnage aléatoire des généalogies, certaines (rares) vont contribuer à la vraisemblance, mais un grand nombre risque d'être improbable. Cela justifie de diriger la méthode MC vers les généalogies qui ont le plus de chance d'être réalisées.

Simulation de la loi d'importance

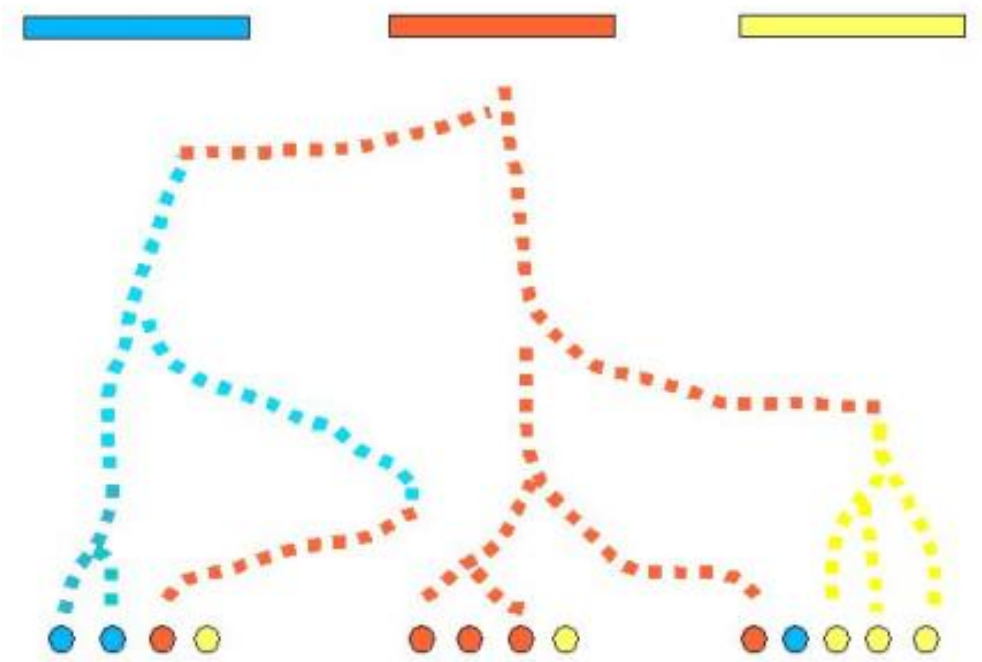
- Les généalogies G sont simulées selon la loi d'importance par McMC

$$q(G) = p(G/D;\theta_0) \propto p(D/G)p(G/\theta_0).$$

- On fait évoluer de manière itérative un arbre initial en ne modifiant qu'un noeud interne à chaque transition et en recalculant les temps de coalescence conditionnels.
- *Méthode de rejet* fondée sur le calcul de $p(D/G)$ (programmation dynamique).

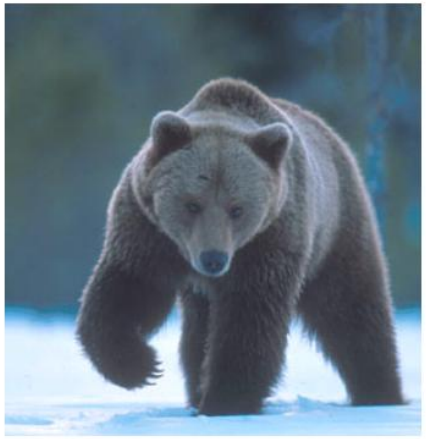
Coalescent Structuré

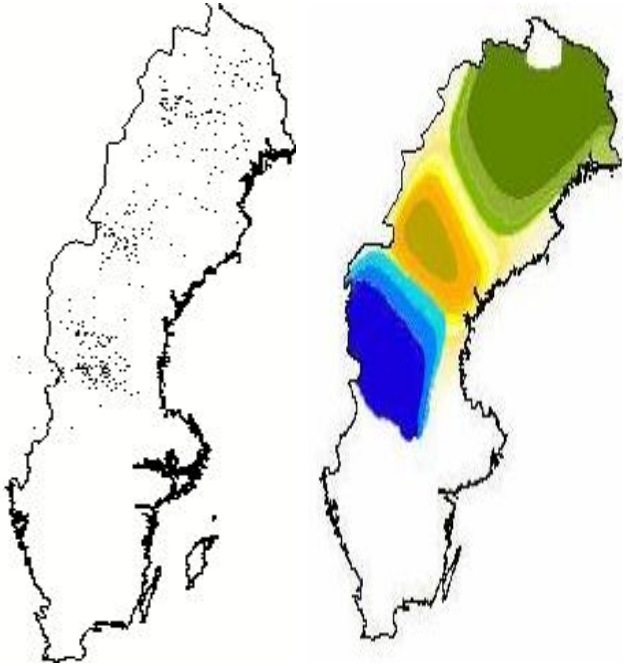
- Généalogie avec des flux de gènes entre populations isolées



Structure génétique de la population

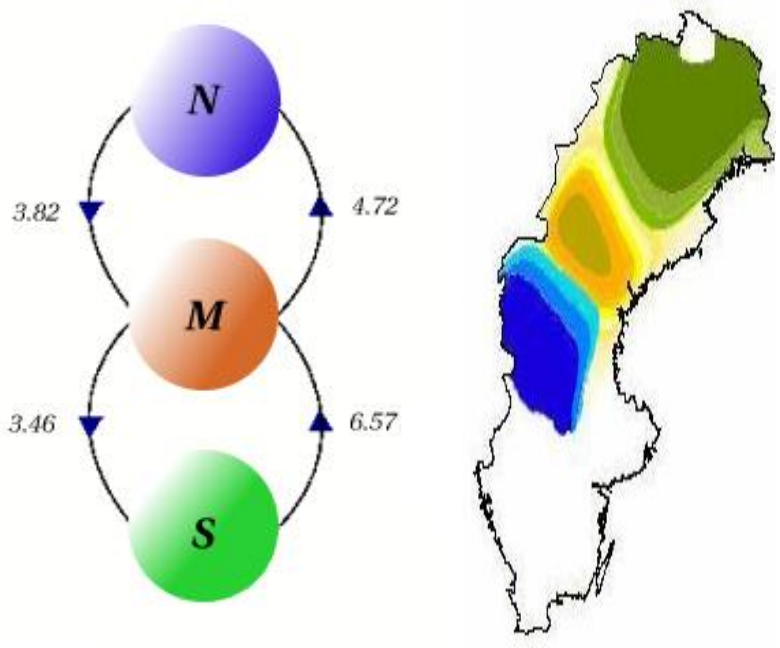
- Méthode de phylogénie fondée sur le modèle de mutation IAM (infinité d'allèles)





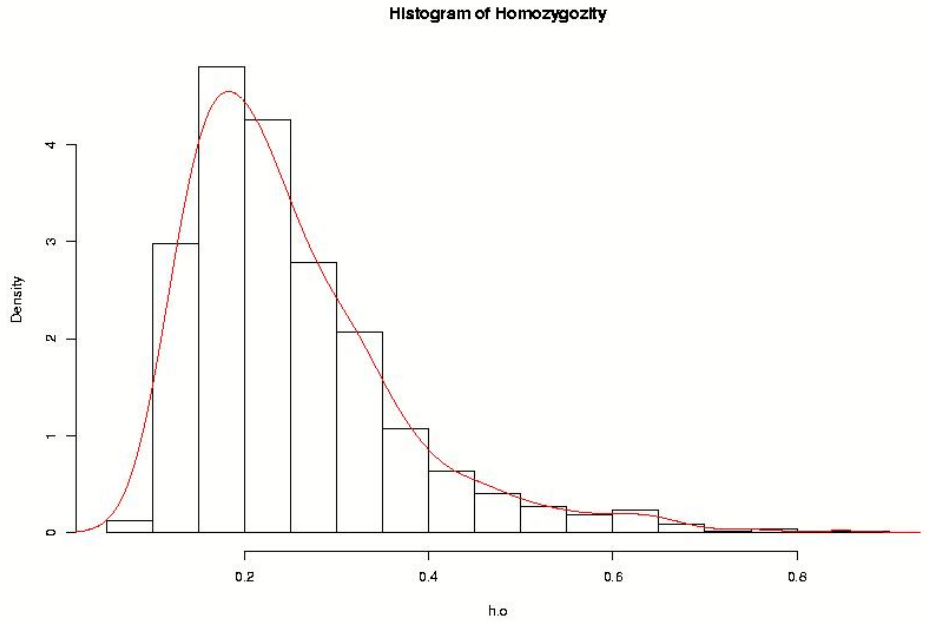
Structure génétique de la population

- Estimation MCMC des flux de gènes : `migrate` (Beerli et al., PNAS 2001)



Structure génétique de la population

- Validation du modèle : Trois populations isolées par la distance et dont les flux de gènes sont véhiculés par les mâles. Bonne compatibilité avec le modèle de mutation IAM.



Estimation de la dispersion spatiale chez Ursus arctos

- Développement d'une méthode de pseudovraisemblance (calcul rapide)
- Utilisation de grappes de PCs pour la correction du biais de la méthode (simulations)
- Estimation du paramètre de dispersion mère/fille $\sigma \approx 8 - 11$ km

Conclusions

- Méthodes destinées aux problèmes de calcul en grandes dimensions
- pour consommateurs voraces en ressources informatiques
- mais nécessaires devant la complexité de certains modèles
- et assez faciles à programmer...