La Méthode de Monte Carlo

Etienne Pardoux

UMR 6632 Laboratoire d'Analyse, Topologie, Probabilités
et EA 3781 Evolution Biologique
Université de Provence

Contents

Outline

Introduction

- La méthode de Monte Carlo est une méthode numérique, qui utilise des tirages aléatoires pour réaliser le calcul d'une quantité déterministe.
- On verra deux applications de cette méthode pour
  1. le calcul du prix d'une option en Finance ;
  2. la détermination d'un arbre phylogénétique à partir de données génomiques ;

  et on expliquera les avantages de celle–ci.
- Mais nous allons tout d'abord présenter les concepts de base du Calcul des Probabilités sur l'exemple le plus simple : une suite de tirages à PILE ou FACE, et le mouvement brownien.

Introduction

- La méthode de Monte Carlo est une méthode numérique, qui utilise des tirages aléatoires pour réaliser le calcul d'une quantité déterministe.
- On verra deux applications de cette méthode pour
  1. le calcul du prix d'une option en Finance ;
  2. la détermination d'un arbre phylogénétique à partir de données génomiques ;

  et on expliquera les avantages de celle–ci.
- Mais nous allons tout d'abord présenter les concepts de base du Calcul des Probabilités sur l'exemple le plus simple : une suite de tirages à PILE ou FACE, et le mouvement brownien.

Introduction

- La méthode de Monte Carlo est une méthode numérique, qui utilise des tirages aléatoires pour réaliser le calcul d'une quantité déterministe.
- On verra deux applications de cette méthode pour
  1. le calcul du prix d'une option en Finance ;
  2. la détermination d'un arbre phylogénétique à partir de données génomiques ;

  et on expliquera les avantages de celle–ci.
- Mais nous allons tout d'abord présenter les concepts de base du Calcul des Probabilités sur l'exemple le plus simple : une suite de tirages à PILE ou FACE, et le mouvement brownien.

Outline

Formalisation du jeu de PILE ou FACE

- Supposons que l'on joue n fois de suite à PILE ou FACE avec une pièce telle que

$$\text{Proba(PILE)} = p, \quad \text{Proba(FACE)} = 1 - p,$$

avec un certain $0 < p < 1$.

- On formalise cette expérience aléatoire en définissant comme suit le résultat du k–ième jet

$$X_k = \begin{cases} 1, & \text{si on obtient PILE au } k\text{–ième coup;} \\ 0, & \text{si l'on obtient FACE au } k\text{–ième coup.} \end{cases}$$

- La LOI DES GRANDS NOMBRES du Calcul des Probabilités nous enseigne que quand n est grand, la proportion de PILE obtenus est proche de p, soit

$$\frac{X_1 + \cdots + X_n}{n} \simeq p.$$

Donc la proportion de PILE obtenus en n coups peut être utilisé pour estimer p, que nous allons considérer comme INCONNU.

Formalisation du jeu de PILE ou FACE

- Supposons que l'on joue n fois de suite à PILE ou FACE avec une pièce telle que

$$\text{Proba(PILE)} = p, \quad \text{Proba(FACE)} = 1 - p,$$

avec un certain $0 < p < 1$.

- On formalise cette expérience aléatoire en définissant comme suit le résultat du k–ième jet

$$X_k = \begin{cases} 1, & \text{si on obtient PILE au } k\text{–ième coup;} \\ 0, & \text{si l'on obtient FACE au } k\text{–ième coup.} \end{cases}$$

- La LOI DES GRANDS NOMBRES du Calcul des Probabilités nous enseigne que quand n est grand, la proportion de PILE obtenus est proche de p, soit

$$\frac{X_1 + \cdots + X_n}{n} \simeq p.$$

Donc la proportion de PILE obtenus en n coups peut être utilisé pour estimer p, que nous allons considérer comme INCONNU.

Formalisation du jeu de PILE ou FACE

- Supposons que l'on joue n fois de suite à PILE ou FACE avec une pièce telle que

$$\text{Proba(PILE)} = p, \quad \text{Proba(FACE)} = 1 - p,$$

avec un certain $0 < p < 1$.

- On formalise cette expérience aléatoire en définissant comme suit le résultat du k–ième jet

$$X_k = \begin{cases} 1, & \text{si on obtient PILE au } k\text{–ième coup;} \\ 0, & \text{si l'on obtient FACE au } k\text{–ième coup.} \end{cases}$$

- La LOI DES GRANDS NOMBRES du Calcul des Probabilités nous enseigne que quand n est grand, la proportion de PILE obtenus est proche de p, soit

$$\frac{X_1 + \cdots + X_n}{n} \simeq p.$$

Donc la proportion de PILE obtenus en n coups peut être utilisé pour estimer p, que nous allons considérer comme INCONNU.

Loi des Grands Nombres

- Un calcul classique donne

$$\begin{aligned}
\text{Esp}\left[\left(\frac{X_1+\cdots+X_n}{n}-p\right)^2\right] &= \text{Esp}\left[\left(\frac{X_1-p}{n}+\cdots+\frac{X_n-p}{n}\right)^2\right] \\
\text{(par l'indépendance des } X_k\text{)} \quad &= \text{Esp}\left[\left(\frac{X_1-p}{n}\right)^2+\cdots+\left(\frac{X_n-p}{n}\right)^2\right] \\
&= \frac{n}{n^2}\text{Esp}[(X_1-p)^2] \\
&= \frac{p(1-p)}{n} \\
&\to 0, \quad \text{quand } n\to\infty
\end{aligned}$$

- Détail du calcul de $\text{Var}(X_1)=\text{Esp}[(X_1-p)^2]$:

$$\text{Var}(X_1)=p\times(1-p)^2+(1-p)\times p^2=p-p^2.$$

- On a en fait un résultat plus précis.

Loi des Grands Nombres

- Un calcul classique donne

$$\begin{aligned}
\text{Esp}\left[\left(\frac{X_1+\cdots+X_n}{n}-p\right)^2\right] &= \text{Esp}\left[\left(\frac{X_1-p}{n}+\cdots+\frac{X_n-p}{n}\right)^2\right] \\
\text{(par l'indépendance des } X_k\text{)} \quad &= \text{Esp}\left[\left(\frac{X_1-p}{n}\right)^2+\cdots+\left(\frac{X_n-p}{n}\right)^2\right] \\
&= \frac{n}{n^2}\text{Esp}[(X_1-p)^2] \\
&= \frac{p(1-p)}{n} \\
&\to 0, \quad \text{quand } n\to\infty
\end{aligned}$$

- Détail du calcul de $\text{Var}(X_1) = \text{Esp}[(X_1-p)^2]$:

$$\text{Var}(X_1) = p\times(1-p)^2+(1-p)\times p^2 = p-p^2.$$

- On a en fait un résultat plus précis.

Loi des Grands Nombres

- Un calcul classique donne

$$\begin{aligned}
\text{Esp}\left[\left(\frac{X_1+\cdots+X_n}{n}-p\right)^2\right] &= \text{Esp}\left[\left(\frac{X_1-p}{n}+\cdots+\frac{X_n-p}{n}\right)^2\right] \\
\text{(par l'indépendance des } X_k\text{)} \quad &= \text{Esp}\left[\left(\frac{X_1-p}{n}\right)^2+\cdots+\left(\frac{X_n-p}{n}\right)^2\right] \\
&= \frac{n}{n^2}\text{Esp}[(X_1-p)^2] \\
&= \frac{p(1-p)}{n} \\
&\to 0, \quad \text{quand } n\to\infty
\end{aligned}$$

- Détail du calcul de $\text{Var}(X_1) = \text{Esp}[(X_1-p)^2]$:

$$\text{Var}(X_1) = p\times(1-p)^2+(1-p)\times p^2 = p-p^2.$$

- On a en fait un résultat plus précis.

Théorème de de Moivre

- Le Théorème de de Moivre (1718) nous dit que

$$\text{Proba}\left(a < \sqrt{\frac{n}{p(1-p}}\left(\frac{X_1+\cdots+X_n}{n}-p\right) < b\right) \to \frac{1}{\sqrt{2\pi}}\int_a^b e^{-x^2/2}dx.$$

- En particulier,

$$\text{Proba}\left(\left|\frac{X_1+\cdots+X_n}{n}-p\right| > c\sqrt{\frac{p(1-p)}{n}}\right) \simeq \frac{2}{\sqrt{2\pi}}\int_c^\infty e^{-x^2/2}dx.$$

Notons que le membre de droite vaut $0,05$ si $c = 1,96$, et $0,01$ si $c = 2,6$.

Théorème de de Moivre

- Le Théorème de de Moivre (1718) nous dit que

$$\text{Proba}\left(a < \sqrt{\frac{n}{p(1-p}}\left(\frac{X_1+\cdots+X_n}{n}-p\right) < b\right) \to \frac{1}{\sqrt{2\pi}}\int_a^b e^{-x^2/2}dx.$$

- En particulier,

$$\text{Proba}\left(\left|\frac{X_1+\cdots+X_n}{n}-p\right| > c\sqrt{\frac{p(1-p)}{n}}\right) \simeq \frac{2}{\sqrt{2\pi}}\int_c^\infty e^{-x^2/2}dx.$$

Notons que le membre de droite vaut $0,05$ si $c = 1,96$, et $0,01$ si $c = 2,6$.

Intervalle de confiance

- Mais p est INCONNU ! Cependant on sait que pour $0 < p < 1$, $p(1-p) \leq 1/4$, donc $\sqrt{p(1-p)} \leq 1/2$. D'où l'on déduit (en choisissant par exemple $c = 1,96$
- $$\text{Proba}\left(\left|\frac{X_1 + \cdots + X_n}{n} - p\right| > \frac{0,98}{\sqrt{n}}\right) \leq 0,05,$$

  où on a un petit peu triché, mais c'est essentiellement vrai dès que $np > 10$.
- Autrement dit,

  $$\text{Proba}\left(p \in \left[\frac{X_1 + \cdots + X_n}{n} - \frac{0,98}{\sqrt{n}}, \frac{X_1 + \cdots + X_n}{n} + \frac{0,98}{\sqrt{n}}\right]\right) \geq 0,95.$$

  Il n'y a aucun intervalle borné de $\mathbb{R}$ dont on puisse dire avec certitude qu'il contient p, mais il y a des intervalles dits *intervalle de confiance*, dont on peut dire qu'il contiennent p avec une probabilité très proche de 1.

Intervalle de confiance

- Mais p est INCONNU ! Cependant on sait que pour $0 < p < 1$, $p(1-p) \leq 1/4$, donc $\sqrt{p(1-p)} \leq 1/2$. D'où l'on déduit (en choisissant par exemple $c = 1,96$
- $$\text{Proba}\left(\left|\frac{X_1 + \cdots + X_n}{n} - p\right| > \frac{0,98}{\sqrt{n}}\right) \leq 0,05,$$

  où on a un petit peu triché, mais c'est essentiellement vrai dès que $np > 10$.
- Autrement dit,

$$\text{Proba}\left(p \in \left[\frac{X_1 + \cdots + X_n}{n} - \frac{0,98}{\sqrt{n}}, \frac{X_1 + \cdots + X_n}{n} + \frac{0,98}{\sqrt{n}}\right]\right) \geq 0,95.$$

Il n'y a aucun intervalle borné de $\mathbb{R}$ dont on puisse dire avec certitude qu'il contient p, mais il y a des intervalles dits *intervalle de confiance*, dont on peut dire qu'il contiennent p avec une probabilité très proche de 1.

Intervalle de confiance

- Mais p est INCONNU ! Cependant on sait que pour $0 < p < 1$, $p(1-p) \leq 1/4$, donc $\sqrt{p(1-p)} \leq 1/2$. D'où l'on déduit (en choisissant par exemple $c = 1,96$
- $$\text{Proba}\left(\left|\frac{X_1 + \cdots + X_n}{n} - p\right| > \frac{0,98}{\sqrt{n}}\right) \leq 0,05,$$

  où on a un petit peu triché, mais c'est essentiellement vrai dès que $np > 10$.
- Autrement dit,

$$\text{Proba}\left(p \in \left[\frac{X_1 + \cdots + X_n}{n} - \frac{0,98}{\sqrt{n}}, \frac{X_1 + \cdots + X_n}{n} + \frac{0,98}{\sqrt{n}}\right]\right) \geq 0,95.$$

Il n'y a aucun intervalle borné de $\mathbb{R}$ dont on puisse dire avec certitude qu'il contient p, mais il y a des intervalles dits *intervalle de confiance*, dont on peut dire qu'il contiennent p avec une probabilité très proche de 1.

Outline

Mouvement brownien

- Il s'agit d'une processus stochastique (i. e. une fonction aléatoire du temps) qui a été introduit par le botaniste R. Brown au début du 19ème siècle pour modéliser les mouvements de grains de pollen en suspension, étudié ensuite au 20ème siècle par A. Einstein, M. Smoluchowsky, N. Wiener, P. Lévy, ...
- La particularité du mouvement brownien est que ses accroissements sont indépendants, ce qui rend ses trajectoires très irrégulières (sa vitesse est infinie en tout point).
- Le mouvement brownien est utilisé dans de nombreux modèles de phénomènes physiques. Il permet aussi de donner une expression de la solution de l'équation de la chaleur.

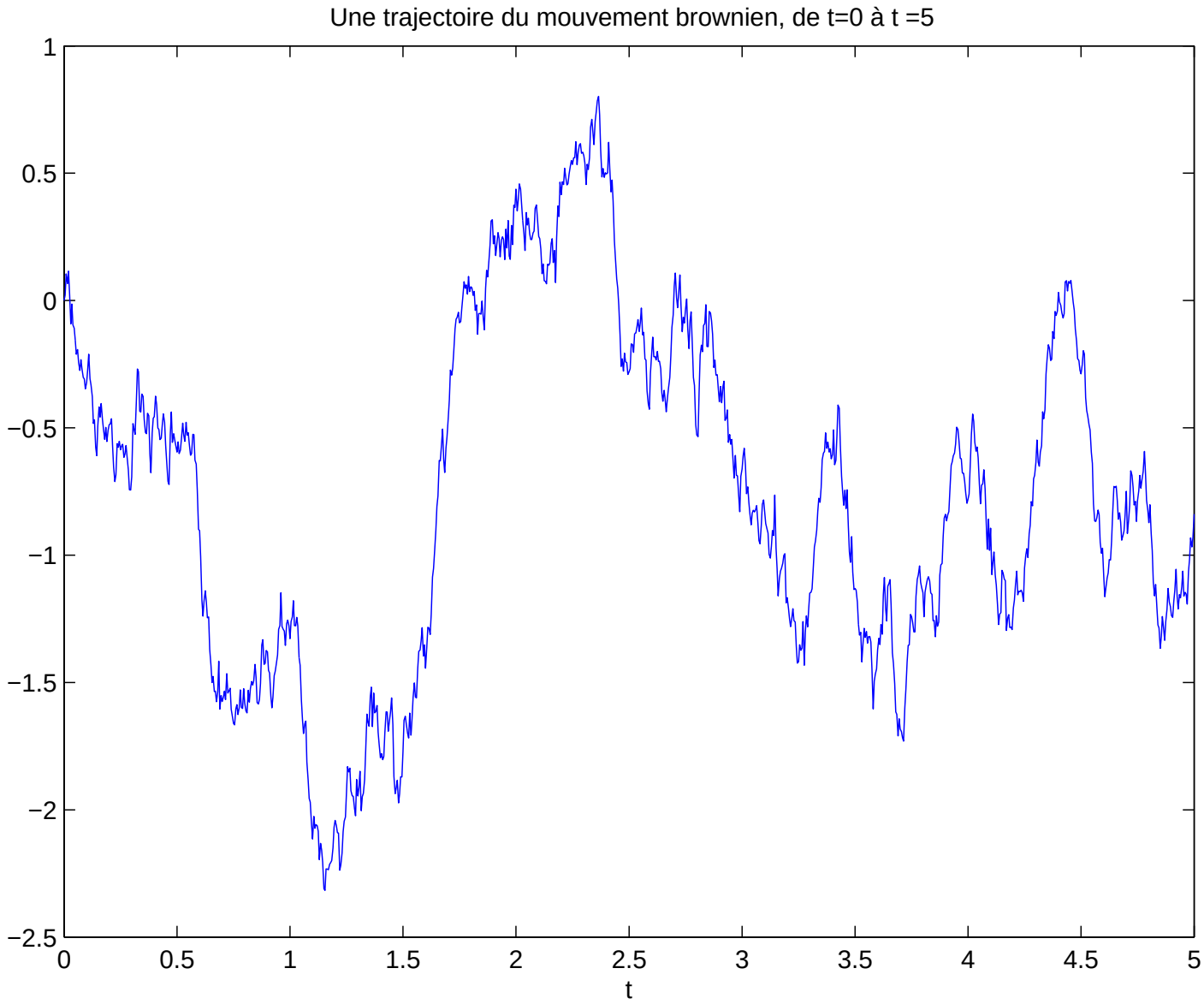
Mouvement brownien

- Il s'agit d'une processus stochastique (i. e. une fonction aléatoire du temps) qui a été introduit par le botaniste R. Brown au début du 19ème siècle pour modéliser les mouvements de grains de pollen en suspension, étudié ensuite au 20ème siècle par A. Einstein, M. Smoluchowsky, N. Wiener, P. Lévy, ...
- La particularité du mouvement brownien est que ses accroissements sont indépendants, ce qui rend ses trajectoires très irrégulières (sa vitesse est infinie en tout point).
- Le mouvement brownien est utilisé dans de nombreux modèles de phénomènes physiques. Il permet aussi de donner une expression de la solution de l'équation de la chaleur.

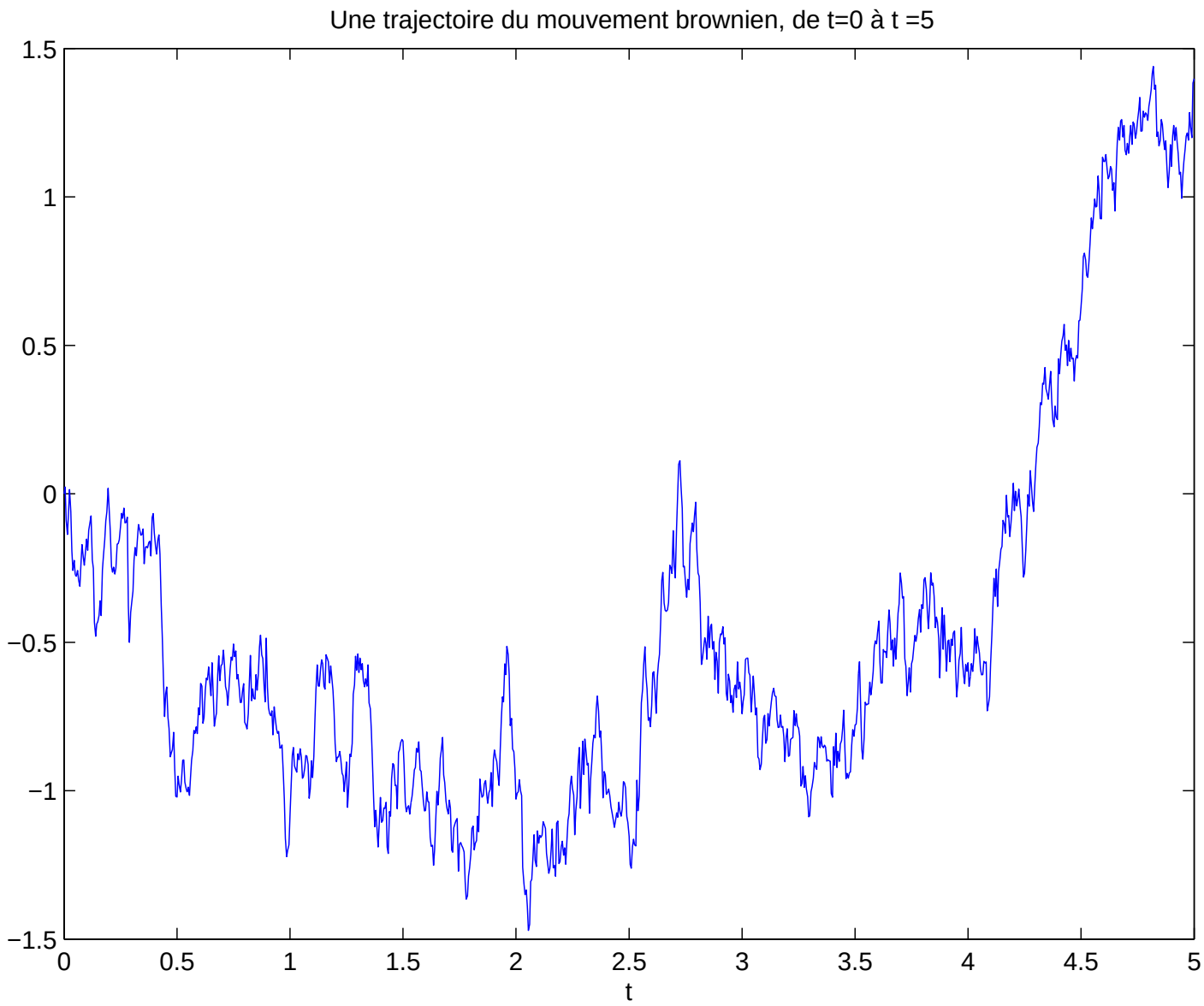
Mouvement brownien

- Il s'agit d'une processus stochastique (i. e. une fonction aléatoire du temps) qui a été introduit par le botaniste R. Brown au début du 19ème siècle pour modéliser les mouvements de grains de pollen en suspension, étudié ensuite au 20ème siècle par A. Einstein, M. Smoluchowsky, N. Wiener, P. Lévy, ...
- La particularité du mouvement brownien est que ses accroissements sont indépendants, ce qui rend ses trajectoires très irrégulières (sa vitesse est infinie en tout point).
- Le mouvement brownien est utilisé dans de nombreux modèles de phénomènes physiques. Il permet aussi de donner une expression de la solution de l'équation de la chaleur.

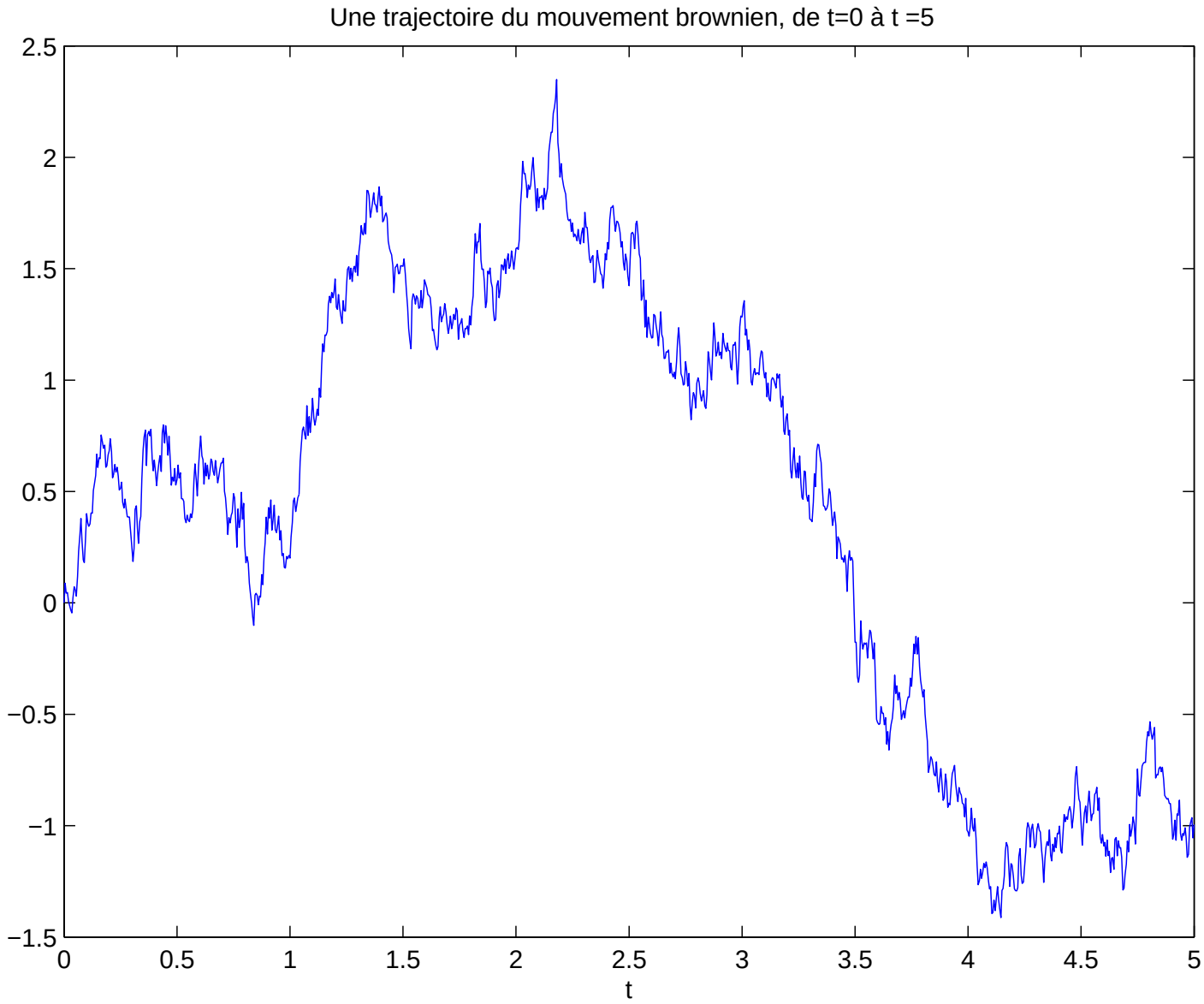
Une trajectoire typique du mouvement brownien 1



Une trajectoire typique du mouvement brownien 2



Une trajectoire typique du mouvement brownien 3



Outline

Le modèle de Black–Scholes

Supposons qu'il existe deux actifs accessibles sur le marché :

- un actif *non risqué* (ex. placement à la Caisse d'Epargne), dont le prix à l'instant t est

$$S_0(t) = S_0(0)\exp(rt);$$

où r est le taux d'intérêt ;

- un actif *risqué* (ex. une action en bourse), dont le prix à l'instant t est

$$S(t) = S(0)\exp(\lambda t + \sigma B(t)),$$

où λ est un paramètre de tendance, σ la *volatilité*, $\{B(t),\ t \geq 0\}$ est un mouvement brownien.

Le modèle de Black–Scholes

Supposons qu'il existe deux actifs accessibles sur le marché :

- un actif *non risqué* (ex. placement à la Caisse d'Epargne), dont le prix à l'instant t est

$$S_0(t) = S_0(0)\exp(rt);$$

où r est le taux d'intérêt ;

- un actif *risqué* (ex. une action en bourse), dont le prix à l'instant t est

$$S(t) = S(0)\exp(\lambda t + \sigma B(t)),$$

où λ est un paramètre de tendance, σ la *volatilité*, $\{B(t),\ t \geq 0\}$ est un mouvement brownien.

Option

- Une option est un contrat qui donne le droit (pas l'obligation) d'acheter (ou de vendre) à l'échéance T n unités de l'actif risqué, à un prix K fixé à l'avance.
- Le but de ce contrat est de couvrir son acquéreur vis à vis de fluctuations à la hausse (ou à la baisse) de cet actif risqué, entre l'instant présent $t = 0$ et l'échéance $t = T$.
- L'acquéreur de cette option doit payer une "prime" au moment de la signature du contrat.

Option

- Une option est un contrat qui donne le droit (pas l'obligation) d'acheter (ou de vendre) à l'échéance T n unités de l'actif risqué, à un prix K fixé à l'avance.
- Le but de ce contrat est de couvrir son acquéreur vis à vis de fluctuations à la hausse (ou à la baisse) de cet actif risqué, entre l'instant présent $t = 0$ et l'échéance $t = T$.
- L'acquéreur de cette option doit payer une "prime" au moment de la signature du contrat.

Option

- Une option est un contrat qui donne le droit (pas l'obligation) d'acheter (ou de vendre) à l'échéance T n unités de l'actif risqué, à un prix K fixé à l'avance.
- Le but de ce contrat est de couvrir son acquéreur vis à vis de fluctuations à la hausse (ou à la baisse) de cet actif risqué, entre l'instant présent $t = 0$ et l'échéance $t = T$.
- L'acquéreur de cette option doit payer une "prime" au moment de la signature du contrat.

Formule de Black–Scholes

Le *juste prix de l'option* (par unité d'actif à acheter ou à vendre) est donné par la célèbre formule de Black–Scholes (1972)

- Dans le cas d'une option d'achat (*call* en Anglais)

$$C = \text{Esp}\left[\left(e^{[\sigma B(T) - \frac{\sigma^2}{2}T]} - e^{-rT}K\right)_+\right].$$

- Dans le cas d'une option de vente (*put* en Anglais)

$$P = \text{Esp}\left[\left(e^{-rT}K - e^{[\sigma B(T) - \frac{\sigma^2}{2}T]}\right)_+\right].$$

- On va considérer ci–dessous le cas d'une option de vente.

Formule de Black–Scholes

Le *juste prix de l'option* (par unité d'actif à acheter ou à vendre) est donné par la célèbre formule de Black–Scholes (1972)

- Dans le cas d'une option d'achat (*call* en Anglais)

$$C = \text{Esp}\left[\left(e^{[\sigma B(T) - \frac{\sigma^2}{2}T]} - e^{-rT}K\right)_+\right].$$

- Dans le cas d'une option de vente (*put* en Anglais)

$$P = \text{Esp}\left[\left(e^{-rT}K - e^{[\sigma B(T) - \frac{\sigma^2}{2}T]}\right)_+\right].$$

- On va considérer ci–dessous le cas d'une option de vente.

Formule de Black–Scholes

Le *juste prix de l'option* (par unité d'actif à acheter ou à vendre) est donné par la célèbre formule de Black–Scholes (1972)

- Dans le cas d'une option d'achat (*call* en Anglais)

$$C = \text{Esp}\left[\left(e^{[\sigma B(T) - \frac{\sigma^2}{2}T]} - e^{-rT}K\right)_+\right].$$

- Dans le cas d'une option de vente (*put* en Anglais)

$$P = \text{Esp}\left[\left(e^{-rT}K - e^{[\sigma B(T) - \frac{\sigma^2}{2}T]}\right)_+\right].$$

- On va considérer ci–dessous le cas d'une option de vente.

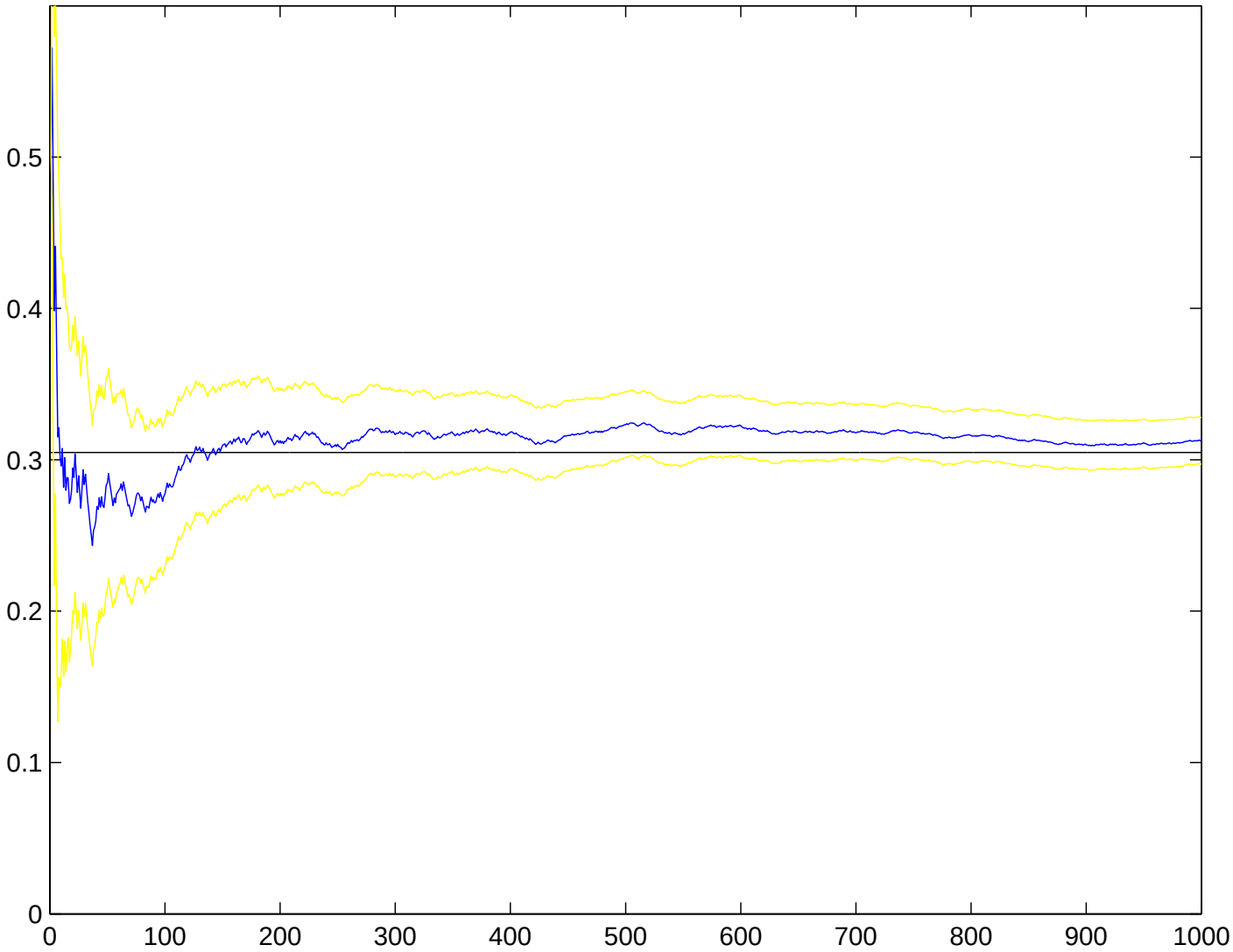
Calcul du prix d'une option de vente par Monte Carlo

- Dans le cas ci–dessus, on a une *formule explicite* pour la valeur P d'une option de vente, en fonction des paramètres r, σ, T et K.
- Dans la figure ci–dessous, on compare la valeur *exacte* avec le résultat du calcul de Monte Carlo, et les bornes de l'intervalle de confiance, en fonction du nombre N de simulations.

Calcul du prix d'une option de vente par Monte Carlo

- Dans le cas ci–dessus, on a une *formule explicite* pour la valeur P d'une option de vente, en fonction des paramètres r, σ, T et K.
- Dans la figure ci–dessous, on compare la valeur *exacte* avec le résultat du calcul de Monte Carlo, et les bornes de l'intervalle de confiance, en fonction du nombre N de simulations.

Résultat numérique



Option panier

- Supposons cette fois que l'on dispose sur le marché de 10 actifs risqués, chacun fluctuant en fonction d'un mouvement brownien différent, avec une volatilité propre.
- Considérons une option de vente sur un panier de ces 10 actions, qui consiste en un contrat aux termes duquel le vendeur de l'action s'engage à vous acheter à l'échéance T au prix fixé K a_i unités de l'action numéro i, de $i = 1$ à $i = 10$.
- La formule de Black–Scholes donne comme prix

$$P = \mathrm{Esp}\left[\left(e^{-rT}K - \sum_{i=1}^{10} a_i e^{[\sigma_i B_i(T) - \frac{\sigma_i^2}{2}T]}\right)_+\right].$$

Option panier

- Supposons cette fois que l'on dispose sur le marché de 10 actifs risqués, chacun fluctuant en fonction d'un mouvement brownien différent, avec une volatilité propre.
- Considérons une option de vente sur un panier de ces 10 actions, qui consiste en un contrat aux termes duquel le vendeur de l'action s'engage à vous acheter à l'échéance T au prix fixé K a_i unités de l'action numéro i, de $i = 1$ à $i = 10$.
- La formule de Black–Scholes donne comme prix

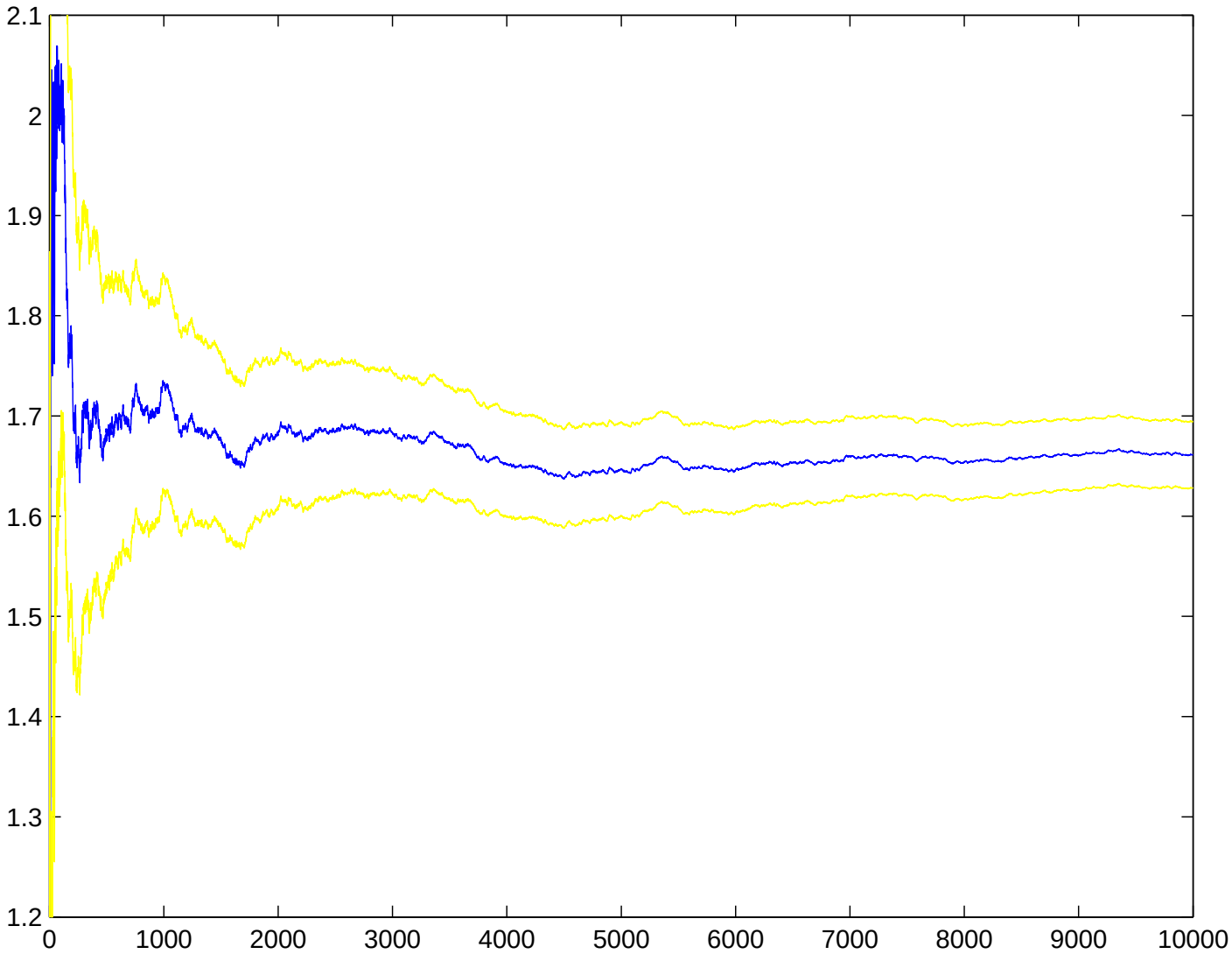
$$P = \mathrm{Esp}\left[\left(e^{-rT}K - \sum_{i=1}^{10} a_i e^{[\sigma_i B_i(T) - \frac{\sigma_i^2}{2}T]}\right)_+\right].$$

Option panier

- Supposons cette fois que l'on dispose sur le marché de 10 actifs risqués, chacun fluctuant en fonction d'un mouvement brownien différent, avec une volatilité propre.
- Considérons une option de vente sur un panier de ces 10 actions, qui consiste en un contrat aux termes duquel le vendeur de l'action s'engage à vous acheter à l'échéance T au prix fixé K a_i unités de l'action numéro i, de $i = 1$ à $i = 10$.
- La formule de Black–Scholes donne comme prix

$$P = \mathrm{Esp}\left[\left(e^{-rT}K - \sum_{i=1}^{10} a_i e^{[\sigma_i B_i(T) - \frac{\sigma_i^2}{2}T]}\right)_+\right].$$

Résultat numérique



Commentaire

- Dans le cas de l'option panier, on n'a pas de formule excate pour le calcul du prix de l'option. Une méthode numérique basée sur une équation aux dérivées partielles est totalement impossible à mettre en oeuvre : il faudrait résoudre une équation en dimension d'espace égale à dix !
- Dans cet exemple, la méthode de Monte Carlo montre toute sa puissance et sa souplesse d'utilisation. Elle est quasiment aussi facile à programmer que la méthode dans le cas d'une seule action, et demande seulement un peu plus de temps de calcul et d'espace mémoire.

Commentaire

- Dans le cas de l'option panier, on n'a pas de formule excate pour le calcul du prix de l'option. Une méthode numérique basée sur une équation aux dérivées partielles est totalement impossible à mettre en oeuvre : il faudrait résoudre une équation en dimension d'espace égale à dix !
- Dans cet exemple, la méthode de Monte Carlo montre toute sa puissance et sa souplesse d'utilisation. Elle est quasiment aussi facile à programmer que la méthode dans le cas d'une seule action, et demande seulement un peu plus de temps de calcul et d'espace mémoire.

Outline

Un problème modèle

- On sait que les trois espèces

  Homme Chimpanzé Gorille

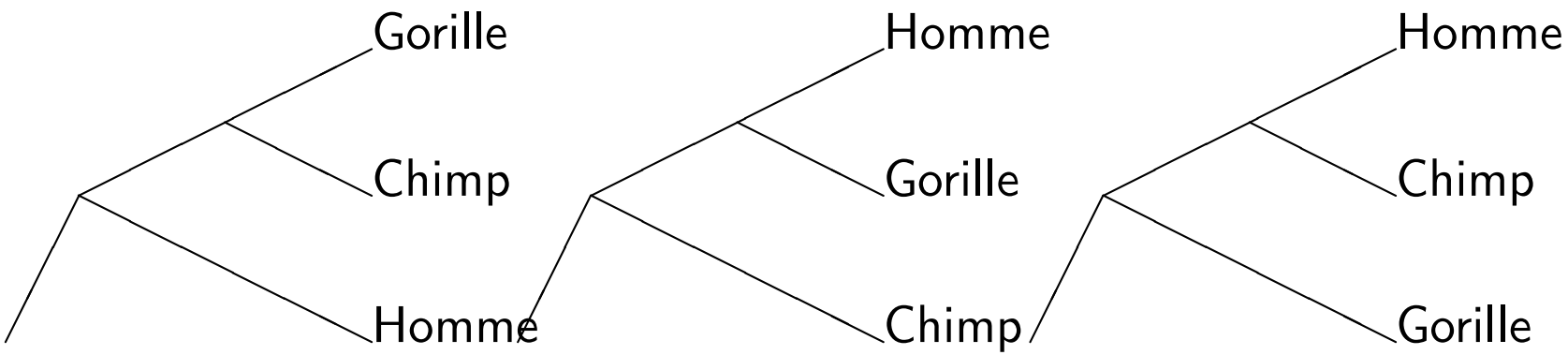
  sont très proches (plus proches entre elles que de touters les autres expèces).
- On voudrait décider laquelle de ces trois formes d'arbre correspond à la réalité de l'histoire des espèces :

Un problème modèle

- On sait que les trois espèces

  Homme Chimpanzé Gorille

  sont très proches (plus proches entre elles que de touters les autres expèces).
- On voudrait décider laquelle de ces trois formes d'arbre correspond à la réalité de l'histoire des espèces :



Les données

- Pour cela, on compare des portions des génomes de ces trois espèces, préalablement alignées.
- Une méthode statistique de plus en plus prisée pour déterminer le bon arbre phylogénétique est la méthode de Bayes, qui consiste à se donner une probabilité a priori sur les trois formes d'arbre possibles, par exemple

$$1/3 \quad 1/3 \quad 1/3,$$

et à calculer, à l'aide de la formule de Bayes une probabilité dite a posteriori, c'est à dire prenant en compte les données.
- Pour $i = 1, 2, 3$,

$$\text{Proba}\,(\text{Arbre } i|\text{Données}) = \frac{\text{Proba}\,(\text{Arbre } i)\,\text{Proba}\,(\text{Données}|\text{Arbre } i)}{\text{Proba}\,(\text{Données})}$$

$$= \frac{\text{Proba}\,(\text{Arbre } i)\,\text{Proba}\,(\text{Données}|\text{Arbre } i)}{\sum_{j=1}^{3}\text{Proba}\,(\text{Arbre } j)\,\text{Proba}\,(\text{Données}|\text{Arbre } j)}$$

Les données

- Pour cela, on compare des portions des génomes de ces trois espèces, préalablement alignées.
- Une méthode statistique de plus en plus prisée pour déterminer le bon arbre phylogénétique est la méthode de Bayes, qui consiste à se donner une probabilité a priori sur les trois formes d'arbre possibles, par exemple

$$1/3 \quad 1/3 \quad 1/3,$$

  et à calculer, à l'aide de la formule de Bayes une probabilité dite a posteriori, c'est à dire prenant en compte les données.
- Pour $i = 1, 2, 3$,

$$\text{Proba}\,(\text{Arbre } i|\text{Données}) = \frac{\text{Proba}\,(\text{Arbre } i)\,\text{Proba}\,(\text{Données}|\text{Arbre } i)}{\text{Proba}\,(\text{Données})}$$

$$= \frac{\text{Proba}\,(\text{Arbre } i)\,\text{Proba}\,(\text{Données}|\text{Arbre } i)}{\sum_{j=1}^{3}\text{Proba}\,(\text{Arbre } j)\,\text{Proba}\,(\text{Données}|\text{Arbre } j)}$$

Les données

- Pour cela, on compare des portions des génomes de ces trois espèces, préalablement alignées.
- Une méthode statistique de plus en plus prisée pour déterminer le bon arbre phylogénétique est la méthode de Bayes, qui consiste à se donner une probabilité a priori sur les trois formes d'arbre possibles, par exemple

$$1/3 \quad 1/3 \quad 1/3,$$

et à calculer, à l'aide de la formule de Bayes une probabilité dite a posteriori, c'est à dire prenant en compte les données.
- Pour $i = 1, 2, 3$,

$$\text{Proba}\,(\text{Arbre } i|\text{Données}) = \frac{\text{Proba}\,(\text{Arbre } i)\,\text{Proba}\,(\text{Données}|\text{Arbre } i)}{\text{Proba}\,(\text{Données})}$$

$$= \frac{\text{Proba}\,(\text{Arbre } i)\,\text{Proba}\,(\text{Données}|\text{Arbre } i)}{\sum_{j=1}^{3} \text{Proba}\,(\text{Arbre } j)\,\text{Proba}\,(\text{Données}|\text{Arbre } j)}$$

Calcul de la formule de Bayes

Dans cette formule,

- Proba(Arbre i) est la probabilité a priori que l'on s'est donnée (ici $(1/3,\ 1/3,\ 1/3)$).
- Proba (Données|Arbre i) est la probabilité d'observer les données, si i est la forme de l'arbre. Cette quantité se calcule à l'aide d'un modèle probabiliste d'évolution des génomes.
- En pratique, la somme qui apparaît au dénominateur est une somme sur un nombre gigantesque de termes ! En effet

Calcul de la formule de Bayes

Dans cette formule,

- Proba(Arbre i) est la probabilité a priori que l'on s'est donnée (ici $(1/3,\ 1/3,\ 1/3)$).
- Proba (Données|Arbre i) est la probabilité d'observer les données, si i est la forme de l'arbre. Cette quantité se calcule à l'aide d'un modèle probabiliste d'évolution des génomes.
- En pratique, la somme qui apparaît au dénominateur est une somme sur un nombre gigantesque de termes ! En effet

Calcul de la formule de Bayes

Dans cette formule,

- Proba(Arbre i) est la probabilité a priori que l'on s'est donnée (ici $(1/3,\ 1/3,\ 1/3)$).
- Proba (Données|Arbre i) est la probabilité d'observer les données, si i est la forme de l'arbre. Cette quantité se calcule à l'aide d'un modèle probabiliste d'évolution des génomes.
- En pratique, la somme qui apparaît au dénominateur est une somme sur un nombre gigantesque de termes ! En effet

Une difficulté

- On traite en général beaucoup plus que 3 espèces, et le nombre d'arbres possibles croît de façon dramatique avec le nombre d'espèces considérées.

Nombre d'espèces	Nombre d'arbres possibles
4	15
5	105
6	945
7	$> 10^4$
10	$> 34 \times 10^6$
20	$8,2 \times 10^{21}$

- Il ne suffit pas de sommer sur les formes possibles d'arbre, il faut aussi prendre en compte les longueurs des branches, et les paramètres du modèle probabiliste d'évolution.

Une difficulté

- On traite en général beaucoup plus que 3 espèces, et le nombre d'arbres possibles croît de façon dramatique avec le nombre d'espèces considérées.

Nombre d'espèces	Nombre d'arbres possibles
4	15
5	105
6	945
7	$> 10^4$
10	$> 34 \times 10^6$
20	$8,2 \times 10^{21}$

- Il ne suffit pas de sommer sur les formes possibles d'arbre, il faut aussi prendre en compte les longueurs des branches, et les paramètres du modèle probabiliste d'évolution.

Comment simuler sous une probabilité impossible à calculer ?

- Pour calculer la probabilité a posteriori par Monte Carlo, il faudrait réaliser des tirages aléatoires sous une probabilité donnée par la formule de Bayes.
- Seul le numérateur de cette formule peut être calculé.
- Autrement dit, on voudrait réaliser des tirages aléatoires sous une probabilité que l'on connaît seulement à une constante de normalisation près, que l'on est incapable de calculer.

Comment simuler sous une probabilité impossible à calculer ?

- Pour calculer la probabilité a posteriori par Monte Carlo, il faudrait réaliser des tirages aléatoires sous une probabilité donnée par la formule de Bayes.
- Seul le numérateur de cette formule peut être calculé.
- Autrement dit, on voudrait réaliser des tirages aléatoires sous une probabilité que l'on connaît seulement à une constante de normalisation près, que l'on est incapable de calculer.

Comment simuler sous une probabilité impossible à calculer ?

- Pour calculer la probabilité a posteriori par Monte Carlo, il faudrait réaliser des tirages aléatoires sous une probabilité donnée par la formule de Bayes.
- Seul le numérateur de cette formule peut être calculé.
- Autrement dit, on voudrait réaliser des tirages aléatoires sous une probabilité que l'on connaît seulement à une constante de normalisation près, que l'on est incapable de calculer.

Monte Carlo par chaîne de Markov

- La solution à ce problème a été inventée par des Physiciens dès 1953, et c'est grâce à cette invention que la statistique bayésienne est utilisable aujourd'hui !
- Une *chaîne de Markov* est une suite de v. a. $\{X_n\}$ de la forme

$$X_{n+1} = f(X_n, Y_n),$$

où les Y_n forment une suite indépendante.
- Sous des hypothèses très faibles, une chaîne de Markov possède une unique probablité invariante π, et la suite $\{X_n\}$ satisfait une loi des grands nombres :

$$\frac{g(X_1) + \cdots + g(X_n)}{n} \to \sum_x g(x)\pi(x)$$

- Enfin, il n'est pas difficile de choisir la bonne chaîne de Markov, si l'on connaît π à une constante multiplicative près.

Monte Carlo par chaîne de Markov

- La solution à ce problème a été inventée par des Physiciens dès 1953, et c'est grâce à cette invention que la statistique bayésienne est utilisable aujourd'hui !
- Une *chaîne de Markov* est une suite de v. a. $\{X_n\}$ de la forme

$$X_{n+1} = f(X_n, Y_n),$$

où les Y_n forment une suite indépendante.

- Sous des hypothèses très faibles, une chaîne de Markov possède une unique probablité invariante π, et la suite $\{X_n\}$ satisfait une loi des grands nombres :

$$\frac{g(X_1) + \cdots + g(X_n)}{n} \to \sum_x g(x)\pi(x)$$

- Enfin, il n'est pas difficile de choisir la bonne chaîne de Markov, si l'on connaît π à une constante multiplicative près.

Monte Carlo par chaîne de Markov

- La solution à ce problème a été inventée par des Physiciens dès 1953, et c'est grâce à cette invention que la statistique bayésienne est utilisable aujourd'hui !
- Une *chaîne de Markov* est une suite de v. a. $\{X_n\}$ de la forme

$$X_{n+1} = f(X_n, Y_n),$$

où les Y_n forment une suite indépendante.

- Sous des hypothèses très faibles, une chaîne de Markov possède une unique probablité invariante π, et la suite $\{X_n\}$ satisfait une loi des grands nombres :

$$\frac{g(X_1) + \cdots + g(X_n)}{n} \to \sum_x g(x)\pi(x)$$

- Enfin, il n'est pas difficile de choisir la bonne chaîne de Markov, si l'on connaît π à une constante multiplicative près.

Monte Carlo par chaîne de Markov

- La solution à ce problème a été inventée par des Physiciens dès 1953, et c'est grâce à cette invention que la statistique bayésienne est utilisable aujourd'hui !
- Une *chaîne de Markov* est une suite de v. a. $\{X_n\}$ de la forme

$$X_{n+1} = f(X_n, Y_n),$$

où les Y_n forment une suite indépendante.

- Sous des hypothèses très faibles, une chaîne de Markov possède une unique probablité invariante π, et la suite $\{X_n\}$ satisfait une loi des grands nombres :

$$\frac{g(X_1) + \cdots + g(X_n)}{n} \to \sum_x g(x)\pi(x)$$

- Enfin, il n'est pas difficile de choisir la bonne chaîne de Markov, si l'on connaît π à une constante multiplicative près.

L'algorithme de Metropolis–Hastings

- Plus précisément, étant donnée une loi de transition markovienne de la forme

$$q(x,y) = \text{Proba}\left(Y_{n+1} = y | Y_n = x\right),$$

- pour passer de X_n à X_{n+1}, on simule d'abord une transition selon $q(x,y)$, et si cette transition consiste à faire passer de x à y, on accepte cette transition avec la probabilité

$$\min\{\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\},$$

(si on la refuse, on pose $X_{n+1} = X_n$).

- Notons que la connaissance des $\pi(x)$ à une constante multiplicative près suffit.

L'algorithme de Metropolis–Hastings

- Plus précisément, étant donnée une loi de transition markovienne de la forme
$$q(x,y) = \text{Proba}\left(Y_{n+1} = y | Y_n = x\right),$$
- pour passer de X_n à X_{n+1}, on simule d'abord une transition selon $q(x,y)$, et si cette transition consiste à faire passer de x à y, on accepte cette transition avec la probabilité
$$\min\{\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\},$$
(si on la refuse, on pose $X_{n+1} = X_n$).
- Notons que la connaissance des $\pi(x)$ à une constante multiplicative près suffit.

L'algorithme de Metropolis–Hastings

- Plus précisément, étant donnée une loi de transition markovienne de la forme
$$q(x,y) = \text{Proba}\left(Y_{n+1} = y | Y_n = x\right),$$
- pour passer de X_n à X_{n+1}, on simule d'abord une transition selon $q(x,y)$, et si cette transition consiste à faire passer de x à y, on accepte cette transition avec la probabilité
$$\min\{\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\},$$
(si on la refuse, on pose $X_{n+1} = X_n$).
- Notons que la connaissance des $\pi(x)$ à une constante multiplicative près suffit.

Avantages et difficultés de cette méthode

- Un gros avantage de la méthode bayésienne de détermination d'un arbre phylogénétique est que l'on obtient non pas un unique arbre, sans savoir quelle confiance on peut attribuer à ce résultat, mais une loi de probabilité sur les arbres.
- Par exemple, dans le cas des 3 espèces *Homme, Gorille, Chimpanzé*, on déduit des données la loi a posteriori

$$\frac{2}{10}, \quad \frac{1}{10}, \quad \frac{7}{10},$$

ce qui indique quelle confiance on peut avoir dans la décision de décider que c'est le 3ème arbre qui correspond à l'histoire des espèces.

- Mais d'un autre coté, on manque de résultats qui donnent des règles précises quand à la façon de mener les calculs, en particulier combien de fois il faut itérer la simulation de la chaîne de Markov X_n.

Avantages et difficultés de cette méthode

- Un gros avantage de la méthode bayésienne de détermination d'un arbre phylogénétique est que l'on obtient non pas un unique arbre, sans savoir quelle confiance on peut attribuer à ce résultat, mais une loi de probabilité sur les arbres.
- Par exemple, dans le cas des 3 espèces *Homme, Gorille, Chimpanzé*, on déduit des données la loi a posteriori

$$\frac{2}{10}, \quad \frac{1}{10}, \quad \frac{7}{10},$$

ce qui indique quelle confiance on peut avoir dans la décision de décider que c'est le 3ème arbre qui correspond à l'histoire des espèces.

- Mais d'un autre coté, on manque de résultats qui donnent des règles précises quand à la façon de mener les calculs, en particulier combien de fois il faut itérer la simulation de la chaîne de Markov X_n.

Avantages et difficultés de cette méthode

- Un gros avantage de la méthode bayésienne de détermination d'un arbre phylogénétique est que l'on obtient non pas un unique arbre, sans savoir quelle confiance on peut attribuer à ce résultat, mais une loi de probabilité sur les arbres.
- Par exemple, dans le cas des 3 espèces *Homme, Gorille, Chimpanzé*, on déduit des données la loi a posteriori

$$\frac{2}{10}, \quad \frac{1}{10}, \quad \frac{7}{10},$$

ce qui indique quelle confiance on peut avoir dans la décision de décider que c'est le 3ème arbre qui correspond à l'histoire des espèces.

- Mais d'un autre coté, on manque de résultats qui donnent des règles précises quand à la façon de mener les calculs, en particulier combien de fois il faut itérer la simulation de la chaîne de Markov X_n.

Conclusion

- Autrement dit, les diverses sciences (en particulier la Biologie) ont besoin de mathématiciens (vous demain ?) qui fassent avancer la connaissance des algorithmes de type Monte Carlo, en particulier Monte Carlo par chaînes de Markov.
- C'est un exemple de plus du fait que les progrès des ordinateurs engendrent des besoins de progrès en mathématique.
- Merci pour votre attention.

Conclusion

- Autrement dit, les diverses sciences (en particulier la Biologie) ont besoin de mathématiciens (vous demain ?) qui fassent avancer la connaissance des algorithmes de type Monte Carlo, en particulier Monte Carlo par chaînes de Markov.
- C'est un exemple de plus du fait que les progrès des ordinateurs engendrent des besoins de progrès en mathématique.
- Merci pour votre attention.

Conclusion

- Autrement dit, les diverses sciences (en particulier la Biologie) ont besoin de mathématiciens (vous demain ?) qui fassent avancer la connaissance des algorithmes de type Monte Carlo, en particulier Monte Carlo par chaînes de Markov.
- C'est un exemple de plus du fait que les progrès des ordinateurs engendrent des besoins de progrès en mathématique.
- Merci pour votre attention.