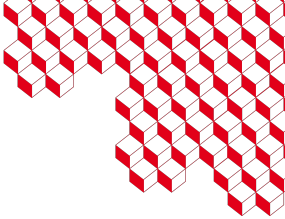

# Méthodes de Monte-Carlo

M2 Radiophysique médicale, INSTN, 2023

Clément GAUCHY (clement.gauchy@cea.fr) Blog: clgch.github.io
CEA SACLAY

# Sommaire

# Les origines

- Le principe des méthodes Monte-Carlo est apparu au laboratoire de Los Alamos, à la fin des années 40
- Idée: Simuler la diffusion des neutrons dans un matériau fissile en utilisant de la simulation aléatoire
- Les méthodes MC sont désormais présentes dans tout les domaines impliquant de la simulation numérique: physique, finance, statistique,...

Figure 1: Stanislaw Ulam, mathématicien et fondateur des méthodes Monte-Carlo

# Pourquoi Monte-Carlo ?

Figure 2: Casino de Monte-Carlo, Monaco

L'oncle de Stanislaw Ulam jouait beaucoup au casino de Monte-Carlo, où les jeux de hasard sont rois !

Les premières simulations Monte-Carlo étaient faites "à la main"...

# Génération de nombres aléatoires

Comment simuler l'aléatoire avec un ordinateur ?

# Génération de nombres aléatoires

## Comment simuler l'aléatoire avec un ordinateur ?

On détermine une suite de nombres dans $[0, 1]$ dit pseudo-aléatoires.

# Génération de nombres aléatoires

## Comment simuler l'aléatoire avec un ordinateur ?

On détermine une suite de nombres dans $[0, 1]$ dit pseudo-aléatoires.

Exemple: Générateur congruentiel linéaire

$$z_{k+1} \equiv (az_k + c) \mod m \qquad x_{k+1} = \frac{z_{k+1}}{m-1}$$

On choisit des bons paramètres $a,\ c,\ m$ pour "tromper" les test statistiques et génerer une loi uniforme $\mathcal{U}([0, 1])$. Le premier terme $x_0$ de la suite est appellé *seed*.

# Génération de nombres aléatoires

## Comment simuler l'aléatoire avec un ordinateur ?

On détermine une suite de nombres dans $[0, 1]$ dit pseudo-aléatoires.

Exemple: Générateur congruentiel linéaire

$$z_{k+1} \equiv (az_k + c) \mod m \qquad x_{k+1} = \frac{z_{k+1}}{m-1}$$

On choisit des bons paramètres $a,\ c,\ m$ pour "tromper" les test statistiques et génerer une loi uniforme $\mathcal{U}([0, 1])$. Le premier terme $x_0$ de la suite est appellé *seed*.

La plupart des languages informatique utilisent des algorithmes plus sophistiqués comme Mersenne-Twister

# Génération de réalisations d'une loi de probabilité arbitraire

Soit $X \sim P$, comment générer des réalisations de $X$ à partir d'échantillon de $U \sim \mathcal{U}([0,1])$ ?

# Génération de réalisations d'une loi de probabilité arbitraire

Soit $X \sim P$, comment générer des réalisations de $X$ à partir d'échantillon de $U \sim \mathcal{U}([0,1])$ ?

Soit $F_X = \mathbb{P}(X \leq x)$ la fonction de répartition de $X$.

$$F_X(X) \sim \mathcal{U}([0,1])$$

On utilise alors la propriété $F_X^{-1}(U) \sim P$.

Méthode très efficace si on a une expression simple de $F_X^{-1}$

# Génération de réalisations d'une loi de probabilité arbitraire

## Méthode d'acceptation rejet de Von Neumann

# Génération de réalisations d'une loi de probabilité arbitraire

## Méthode d'acceptation rejet de Von Neumann

On veut échantilloner $X$ de densité de proba $f$ et on sait échantilloner $Y$ de loi $g$ tel que $f \leq M \times g$.

# Génération de réalisations d'une loi de probabilité arbitraire

## Méthode d'acceptation rejet de Von Neumann

On veut échantilloner $X$ de densité de proba $f$ et on sait échantilloner $Y$ de loi $g$ tel que $f \leq M \times g$.

- On simule $U \sim \mathcal{U}([0, 1])$
- On simule $Y \sim g$.
- Si $U < f(Y)/Mg(Y)$, alors on accepte le $Y$ simulé comme un tirage selon $f$

# Exemple: échantillonage du libre parcours d'une particule

Une particule se déplace dans un matériau, sa probabilité d'intéragir entre une distance $x$ et $x + dx$ est

$$\Sigma dx$$

avec $\Sigma$ la section efficace macroscopique (en $m^{-1}$).

# Exemple: échantillonage du libre parcours d'une particule

Une particule se déplace dans un matériau, sa probabilité d'intéragir entre une distance $x$ et $x + dx$ est

$$\Sigma dx$$

avec $\Sigma$ la section efficace macroscopique (en $m^{-1}$).

On note $P(x)$ la probabilité que le particule ait atteint la distance $x$ sans interactions.

$$\begin{aligned} P(x+dx) &= P(x)\mathbb{P}(\text{aucune interactions entre } [x, x+dx]) \quad \texttt{Hypothèse d'indépendance} \\ P(x+dx) &= P(x)(1-\Sigma dx) \\ \frac{dP}{dx} &= -P(x)\Sigma \end{aligned}$$

On a donc $\boxed{P(x) = \exp(-\Sigma x)}$

# Exemple: échantillonage du libre parcours d'une particule

Probabilité de ne pas intéragir jusqu'à la distance puis d'intéragir en $x + dx$:

$$P(x)\Sigma dx = \underbrace{\Sigma \exp(-\Sigma x)}_{\texttt{densité de probabilité}} \quad dx$$

La fonction de répartition $F(x) = \int_0^x \Sigma \exp(-\Sigma s) ds = 1 - \exp(-\Sigma x)$ est facile à inverser !

$$1 - \exp(-\Sigma x) = u \iff x = \frac{-\ln(u)}{\Sigma}$$

# Exemple: échantillonage du libre parcours d'une particule

Soit $X$ la variable aléatoire du libre parcours d'une particule dans le matériau. Elle a pour densité $\Sigma \exp(-\Sigma x)$.

# Exemple: échantillonage du libre parcours d'une particule

Soit $X$ la variable aléatoire du libre parcours d'une particule dans le matériau. Elle a pour densité $\Sigma \exp(-\Sigma x)$.

Le libre parcours moyen $\ell$ est $\ell = \mathbb{E}[X] = \int_0^{+\infty} x\Sigma \exp(-\Sigma x) dx$.

# Exemple: échantillonage du libre parcours d'une particule

Soit $X$ la variable aléatoire du libre parcours d'une particule dans le matériau. Elle a pour densité $\Sigma \exp(-\Sigma x)$.

Le libre parcours moyen $\ell$ est $\ell = \mathbb{E}[X] = \int_0^{+\infty} x\Sigma \exp(-\Sigma x) dx$.

Á partir d'un échantillon $(X_i)_{1 \leq i \leq N}$ i.i.d génerés selon la loi de $X$, on peut utiliser la loi des grands nombres pour faire l'approximation suivante:

$$\ell \approx \frac{1}{N} \sum_{i=1}^{N} X_i$$

# Exemple: échantillonage du libre parcours d'une particule

Soit $X$ la variable aléatoire du libre parcours d'une particule dans le matériau. Elle a pour densité $\Sigma \exp(-\Sigma x)$.

Le libre parcours moyen $\ell$ est $\ell = \mathbb{E}[X] = \int_0^{+\infty} x\Sigma \exp(-\Sigma x) dx$.

Á partir d'un échantillon $(X_i)_{1 \leq i \leq N}$ i.i.d génerés selon la loi de $X$, on peut utiliser la loi des grands nombres pour faire l'approximation suivante:

$$\ell \approx \frac{1}{N} \sum_{i=1}^{N} X_i$$

Les méthodes Monte-Carlo peuvent servir a calculer des intégrales

## Monte-Carlo pour la quadrature numérique

On cherche à calculer $I = \mathbb{E}[g(X)] = \int_{\mathbb{R}^d} g(x)f(x)dx$ avec $f$ la densité de proba. de $X$

# Monte-Carlo pour la quadrature numérique

On cherche à calculer $I = \mathbb{E}[g(X)] = \int_{\mathbb{R}^d} g(x)f(x)dx$ avec $f$ la densité de proba. de $X$

Estimateur Monte-Carlo: la moyenne empirique à partir de $N$ simulations $(X_i)_{1 \leq i \leq N}$ i.i.d. de même loi que $X$.

$$\widehat{I}_N = \frac{1}{N} \sum_{i=1}^{N} g(X_i)$$

# Monte-Carlo pour la quadrature numérique

On cherche à calculer $I = \mathbb{E}[g(X)] = \int_{\mathbb{R}^d} g(x)f(x)dx$ avec $f$ la densité de proba. de $X$

Estimateur Monte-Carlo: la moyenne empirique à partir de $N$ simulations $(X_i)_{1\leq i\leq N}$ i.i.d. de même loi que $X$.

$$\widehat{I}_N = \frac{1}{N}\sum_{i=1}^{N} g(X_i)$$

Propriétés:

- Estimateur sans biais $\mathbb{E}[\widehat{I}_N] = I$
- Convergence (dite "forte") asymptotique grâce à la lois des grands nombres: $\widehat{I}_N \xrightarrow[N\to+\infty]{} I$
- Variance de l'estimateur MC:

$$\mathrm{Var}(\widehat{I}_N) = \frac{1}{N}\mathrm{Var}(g(X))$$

Inconvénient: Convergence lente en $1/\sqrt{N}$
Avantage: Vitesse de convergence indépendante de la dimension $d$ de $X$

## Contrôle de l'erreur d'estimation

On utilise la notion d'intervalle de confiance pour contrôler l'erreur sur $\widehat{I_N}$.

# Contrôle de l'erreur d'estimation

On utilise la notion d'intervalle de confiance pour contrôler l'erreur sur $\widehat{I_N}$.

Estimateur de la variance

$$S_N^2 = \frac{1}{N}\sum_{i=1}^{N}(g(X_i) - \widehat{I_N})^2$$

# Contrôle de l'erreur d'estimation

On utilise la notion d'intervalle de confiance pour contrôler l'erreur sur $\widehat{I_N}$.

Estimateur de la variance

$$S_N^2 = \frac{1}{N}\sum_{i=1}^{N}(g(X_i) - \widehat{I_N})^2$$

Pour $N$ petit:

$$\mathbb{P}(I \in [\widehat{I_N} - t_{N-1,\frac{1+\alpha}{2}}\frac{S_N}{\sqrt{N-1}}, \widehat{I_N} + t_{N-1,\frac{1+\alpha}{2}}\frac{S_N}{\sqrt{N-1}}]) \approx \alpha$$

# Contrôle de l'erreur d'estimation

On utilise la notion d'intervalle de confiance pour contrôler l'erreur sur $\widehat{I_N}$.

Estimateur de la variance

$$S_N^2 = \frac{1}{N}\sum_{i=1}^{N}(g(X_i) - \widehat{I_N})^2$$

Pour $N$ petit:

$$\mathbb{P}(I \in [\widehat{I_N} - t_{N-1,\frac{1+\alpha}{2}}\frac{S_N}{\sqrt{N-1}}, \widehat{I_N} + t_{N-1,\frac{1+\alpha}{2}}\frac{S_N}{\sqrt{N-1}}]) \approx \alpha$$

Pour $N$ grand:

$$\mathbb{P}(I \in [\widehat{I_N} - u_{\frac{1+\alpha}{2}}\frac{S_N}{\sqrt{N-1}}, \widehat{I_N} + u_{\frac{1+\alpha}{2}}\frac{S_N}{\sqrt{N-1}}]) \approx \alpha$$

# Contrôle de l'erreur d'estimation

On utilise la notion d'intervalle de confiance pour contrôler l'erreur sur $\widehat{I_N}$.

Estimateur de la variance

$$S_N^2 = \frac{1}{N}\sum_{i=1}^{N}(g(X_i) - \widehat{I_N})^2$$

Pour $N$ petit:

$$\mathbb{P}(I \in [\widehat{I_N} - t_{N-1,\frac{1+\alpha}{2}}\frac{S_N}{\sqrt{N-1}}, \widehat{I_N} + t_{N-1,\frac{1+\alpha}{2}}\frac{S_N}{\sqrt{N-1}}]) \approx \alpha$$

Pour $N$ grand:

$$\mathbb{P}(I \in [\widehat{I_N} - u_{\frac{1+\alpha}{2}}\frac{S_N}{\sqrt{N-1}}, \widehat{I_N} + u_{\frac{1+\alpha}{2}}\frac{S_N}{\sqrt{N-1}}]) \approx \alpha$$

C'est $\mathrm{Var}(\widehat{I_N})$ qui pilote la largeur de l'intervalle de confiance !

# Réduction de variance

On rappelle que $\mathrm{Var}(\widehat{I_N}) = \frac{1}{N}\mathrm{Var}(g(X))$.

# Réduction de variance

On rappelle que $\mathrm{Var}(\widehat{I_N}) = \frac{1}{N}\mathrm{Var}(g(X))$.

Contrôler la variance de l'intégrande $\iff$ Contrôler la précision de la méthode Monte-Carlo

# Réduction de variance

On rappelle que $\mathrm{Var}(\widehat{I_N}) = \frac{1}{N}\mathrm{Var}(g(X))$.

Contrôler la variance de l'intégrande $\iff$ Contrôler la précision de la méthode Monte-Carlo

Il existe toute une variété de méthodes de réduction de variance:

- Échantillonage d'importance
- Stratification
- Variable de contrôle
- Conditionnement
- ...

# Réduction de variance par variable de contrôle (*control variates*)

- Soit une fonction $h(X)$ appelée variable de contrôle dont on connait l'espérance $\mu = \mathbb{E}(h(X))$
- On définit la variable aléatoire en fonction d'une constante $\alpha$ :

$$W_\alpha(X) = g(X) + \alpha(h(X) - \mu) \rightarrow \mathbb{E}(W_\alpha(X)) = \mathbb{E}(g(X))$$

$$\widehat{I}_\alpha = \frac{1}{N}\sum_{i=1}^{N} W_\alpha(X_i)$$

- Le calcul de l'intégrale peut donc se faire sur la fonction $W_\alpha(X)$. Sa variance est :

$$\mathrm{Var}(W_\alpha(X)) \quad = \quad \mathrm{Var}(g(X)) + \alpha^2\mathrm{Var}(h(X)) + 2\alpha\mathrm{Cov}(g(X), h(X))$$

- Comme fonction de $\alpha$, la variance de $W_\alpha(X)$ atteint son minimun pour la valeur :

$$\begin{aligned}
\alpha_{\mathrm{opt}} &= -\frac{\mathrm{Cov}(g(X), h(X))}{\mathrm{Var}(h(X))} \\
\mathrm{Var}(W_{\alpha_{\mathrm{opt}}}(X)) &= \mathrm{Var}(g(X)) - \underbrace{\frac{[\mathrm{Cov}(g(X), h(X))]^2}{\mathrm{Var}(h(X))}}_{\texttt{réduction de la variance}} \\
&= \mathrm{Var}(g(X))(1 - \rho^2_{g(X),h(X)})
\end{aligned}$$

  en notant $\rho_{g(X),h(X)}$ le coefficient de corrélation entre les variables $g(X)$ et $h(X)$

- Intérêt de choisir une variable de controle la plus corrélée à $g(X)$ (pas toujours évident)

# Exemple d'utilisation d'une variable de contrôle

- Calcul de $I = \int_0^1 g(x)dx$
- $g(x) = 1 + x \rightarrow I = \ln(2)$
- Par tirages MC d'une loi uniforme $X \sim \mathcal{U}(0,1)$ :

$$\bar{G}_n = (1/n) \sum_{i=1}^{n} \frac{1}{1+X_i}$$

- On prend comme variable de contrôle
  $h(X) = 1 + X$ , $\mu = 3/2$
- On peut calculer
  $\rho_{g(X),h(X)} \approx 0.6$

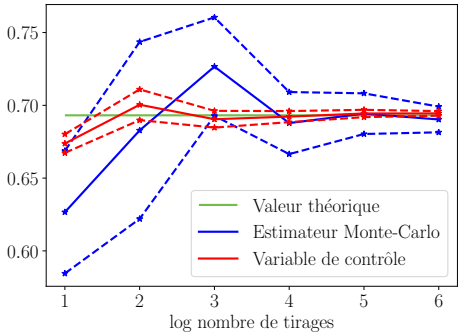


Figure 3: Les courbes en pointillés correspondent à l'intervalle de confiance à 95%

- Dans la figure, les intervalles de confiance asymptotique ont été estimés (TCL : $\bar{G}_n \rightarrow$ loi normale)

# Échantillonage d'importance (*Importance sampling*)

- Calcul de l'intégrale $I = \int_D g(x)f(x)dx$ où $x \in \mathbb{R}^d$ et $g(x)$ une fonction de $D \subset \mathbb{R}^d$ dans $\mathbb{R}$ et $f$ une certaine densité de probabilité
- La représentation de $\mathrm{I}$ comme une espérance n'est pas unique:

$$I = \int_D g(x)f(x)dx = \int_D \frac{g(x)f(x)}{h(x)} h(x)dx = \mathbb{E}_{X\sim h}\left[\frac{g(x)f(x)}{h(x)}\right]$$

- Idée: On peut biaiser l'échantillonage en simulant $X$ selon $g$ pour rendre plus probable les réalisations "importantes".
- On propose l'estimateur suivant:

$$\widehat{I_n} = \frac{1}{N}\sum_{i=1}^{N} g(X_i)\frac{f(X_i)}{h(X_i)}$$

avec $f(x)/h(x)$ appellé le rapport de vraisemblance

# Echantillonage d'importance (2)

- L'estimateur est non biaisé:

$$\mathbb{E}[\widehat{I_N}] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_h\left[g(X_i)\frac{f(X_i)}{h(X_i)}\right] = \mathbb{E}_h\left[g(X)\frac{f(X)}{h(X)}\right] = \int_D \frac{g(x)f(x)}{h(x)}h(x)dx = \mathrm{I}$$

- Convergence de l'estimateur (par la loi forte des grand nombres):

$$\widehat{I_N} \xrightarrow[N\to+\infty]{} I$$

- La variance de l'estimateur s'écrit:

$$\mathrm{Var}(\widehat{I_N}) = \frac{1}{N}\mathrm{Var}_h\left(g(X)\frac{f(X)}{h(X)}\right) = \frac{1}{N}\left(\mathbb{E}_{X\sim f}\left[g(X)^2\frac{f(X)}{h(X)}\right] - I^2\right)$$

Le choix astucieux de $h$ peut réduire drastiquement la variance !

# Échantillonage d'importance optimal

La meilleur distribution de probabilité $h$ est celle minimisant $\mathrm{Var}(\widehat{I_N}) \hookrightarrow$ Soit $h^*$ la meilleur distribution, alors

$$h^* \in \underset{h}{\operatorname{argmin}}\, \mathbb{E}_{X\sim f}\left[g(X)^2 \frac{f(X)}{h(X)}\right] = \int_D g(x)^2 \frac{f^2(x)}{h(x)} dx$$

# Échantillonage d'importance optimal

La meilleur distribution de probabilité $h$ est celle minimisant $\mathrm{Var}(\widehat{I_N}) \hookrightarrow$ Soit $h^*$ la meilleur distribution, alors

$$h^* \in \underset{h}{\operatorname{argmin}}\, \mathbb{E}_{X\sim f}\left[g(X)^2 \frac{f(X)}{h(X)}\right] = \int_D g(x)^2 \frac{f^2(x)}{h(x)} dx$$

La solution de ce problème de minimisation est:

$$h^*(x) = \frac{g(x)f(x)}{\int_D g(u)f(u)du}$$

# Échantillonage d'importance optimal

La meilleur distribution de probabilité $h$ est celle minimisant $\mathrm{Var}(\widehat{I_N}) \hookrightarrow$ Soit $h^*$ la meilleur distribution, alors

$$h^* \in \underset{h}{\operatorname{argmin}}\, \mathbb{E}_{X\sim f}\left[g(X)^2 \frac{f(X)}{h(X)}\right] = \int_D g(x)^2 \frac{f^2(x)}{h(x)} dx$$

La solution de ce problème de minimisation est:

$$h^*(x) = \frac{g(x)f(x)}{\int_D g(u)f(u)du}$$

On peut remarquer que $\mathrm{Var}_{h^*}\left(g(X)\frac{f(X)}{h^*(X)}\right) = 0$ !

# Échantillonage d'importance optimal

La meilleur distribution de probabilité $h$ est celle minimisant $\mathrm{Var}(\widehat{I_N}) \hookrightarrow$ Soit $h^*$ la meilleur distribution, alors

$$h^* \in \underset{h}{\mathrm{argmin}}\, \mathbb{E}_{X\sim f}\left[g(X)^2 \frac{f(X)}{h(X)}\right] = \int_D g(x)^2 \frac{f^2(x)}{h(x)} dx$$

La solution de ce problème de minimisation est:

$$h^*(x) = \frac{g(x)f(x)}{\int_D g(u)f(u)du}$$

On peut remarquer que $\mathrm{Var}_{h^*}\left(g(X)\frac{f(X)}{h^*(X)}\right) = 0$ !

⚠ Le dénominateur de $h^*$ est... $I = \int_D g(x)f(x)dx$ la quantité que l'on cherche à estimer ! Cette loi n'est pas utile en pratique, mais on peut chercher à l'approcher par une famille paramétrique de lois $\{h_\theta, \theta \in \Theta\}$.

$$\theta_* \in \underset{\theta\in\Theta}{\mathrm{argmin}}\, D(h^*, h_\theta)$$

# Exemple de réduction de variance par échantillonnage d'importance

- $g(x) = 3x^2$ et intégrale $I = \int_0^1 g(x)dx = 1$
- Choix uniforme $U(0,1)$, $f(x) = 1_{[0,1]}(x)$

$$\mathrm{Var}(g(X)) \quad = \quad \int_0^1 (3x^2 - 1)^2 dx = \frac{4}{5}$$

- Choix plus astucieux par échantillonnage d'importance : $f(x) = 2x\, 1_{[0,1]}(x)$.

$$\frac{g(x)}{f(x)} \quad = \quad \frac{3x}{2}$$

$$\mathrm{Var}_f\left(\frac{g(X)}{f(X)}\right) \quad = \quad \int_0^1 (\frac{3x}{2} - 1)^2 2x dx = \frac{1}{8}$$

- La variance a été divisé d'un facteur 6

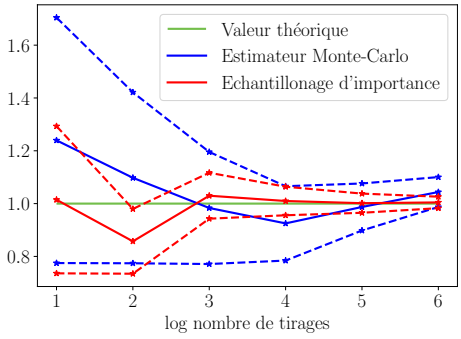


Figure 4: Les courbes en pointillés correspondent à l'intervalle de confiance à 95%

# Sommaire

# Définition d'une chaîne de Markov

- Une chaîne de Markov est un modéle aléatoire pour lequel la probabilité des états du futur ne dépend que de l'état présent
- Soit $(X_t)_{t\geq 0}$ une suite de variables aléatoires à valeurs dans un ensemble $E$ supposé fini $E = \{1, 2, \ldots, M\}$ appelé espace des états
- $(X_t)_{t\geq 0}$ est une chaîne de Markov si pour tout $t \geq 1$ et toute suite $(i_0, i_1, \ldots, i_{t-1}, i, j)$

$$\mathbb{P}(X_{t+1} = j | X_0 = i_0, \ldots, X_{t-1} = i_{t-1}, X_t = i) = \mathbb{P}(X_{t+1} = j | X_t = i)$$

- Autrement dit, le futur est totalement conditionné par le présent car dés qu'on connait le présent (la valeur de $X_t$) la loi du futur $(X_{t+1})$ est parfaitement définie sans connaissance du passé
- Chaîne est dite homogène lorsque la probabilité de transition ne dépend pas de $t$

$$p_{ij} = \mathbb{P}(X_{t+1} = j | X_t = i)$$

- On appelle matrice de transition de la chaîne la matrice $\boldsymbol{P}$ de taille $M \times M$ :

$$\boldsymbol{P} = [p_{ij}]_{1 \leq i,j, \leq M}$$

# Propriétés

- Il est trés facile de calculer la loi jointe de $(X_0, X_1, \ldots, X_t)$ à partir de la loi initiale

$$P(X_0 = i_0, X_1 = i_1, \ldots, X_t = i_t) = \mathbb{P}(X_0 = i_0) p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{t-1} i_t}$$

- La somme par ligne de $\boldsymbol{P}$ est égale à 1. En effet :

$$\begin{aligned}
\sum_{j=1}^{M} p_{ij} &= \sum_{j=1}^{M} \mathbb{P}(X_{t+1} = j | X_t = i) = \sum_{j=1}^{M} \frac{\mathbb{P}(X_{t+1} = j, X_t = i)}{\mathbb{P}(X_t = i)} \\
&= \frac{1}{\mathbb{P}(X_t = i)} \sum_{j=1}^{M} \mathbb{P}(X_{t+1} = j, X_t = i), \quad \texttt{événements disjoints} \\
&= \frac{1}{\mathbb{P}(X_t = i)} \mathbb{P}(X_{t+1} \in \{1, 2, \ldots, M\}, X_t = i) = \frac{\mathbb{P}(X_t = i)}{\mathbb{P}(X_t = i)} = 1
\end{aligned}$$

- La matrice $\boldsymbol{P}$ admet donc $\mathbf{1}$ comme vecteur propre et 1 pour valeur propre associée :

$$\boldsymbol{P}\mathbf{1} = 1 \times \mathbf{1}$$

- Les matrices vérifiant ces propriétés sont appelées matrices stochastiques ou markoviennes

# Equations de Chapman-Kolmogorov

- On note $\pi^t$ la loi de probabilité de la variable $X_t$ définie par le vecteur ligne :

$$\pi^t = (\mathbb{P}(X_t = 1), \mathbb{P}(X_t = 2), \ldots, \mathbb{P}(X_t = M))$$

- A partir des probabilités conditionnelles et en se rappelant que les évènements $\{X_t = j\}$ et $\{X_t = k\}$ sont disjoints si $j \neq k$

$$\begin{aligned}
\mathbb{P}(X_{t+1} = i, X_t = j) &= \mathbb{P}(X_t = j) \times \mathbb{P}(X_{t+1} = i | X_t = j) \\
\sum_j \mathbb{P}(X_{t+1} = i, X_t = j) &= \sum_j \mathbb{P}(X_t = j)\mathbb{P}(X_{t+1} = i | X_t = j) \\
\mathbb{P}(X_{t+1} = i, \bigcup_j X_t = j) &= \sum_j \mathbb{P}(X_t = j)\mathbb{P}(X_{t+1} = i | X_t = j) \\
\{\bigcup_j X_t = j\} = \Omega \rightarrow \mathbb{P}(X_{t+1} = i) &= \sum_j \mathbb{P}(X_t = j)\mathbb{P}(X_{t+1} = i | X_t = j) \\
\pi_i^{t+1} &= \sum_j \pi_j^t p_{ji} \\
\rightarrow \pi^{t+1} &= \pi^t \boldsymbol{P}
\end{aligned}$$

- On en déduit les équations de Chapman-Kolmogorov :

$$\pi^t = \pi^0 \boldsymbol{P}, \quad \mathbb{P}(X_t = j | X_0 = i) = (\boldsymbol{P}^t)_{ij}$$

# Ergodicité, chaînes irréductibles

- Sous certaines conditions sur $P$, la loi de distribution $\pi^t$ tend vers une loi $\pi$ qui devient invariante :

$$\pi = \pi \boldsymbol{P}$$

- Pour chaque matrice de transition il existe au moins une distribution invariante qui peut ne pas être unique.
- Une matrice de transition $\boldsymbol{P}$ est régulière si il existe $t > 0$ tel que la matrice $\boldsymbol{P}^t$ a tous ses éléments strictement positifs. Dans ce cas tous les états sont visités au cours du temps, ce qui correspond à la propriété physique d'ergodicité.
- Pour une chaîne de Markov régulière et donc ergodique, la distribution stationnaire $\boldsymbol{\mu}$ est unique
- La propriété d'ergodicité garantit alors la convergence des $X_t$ vers une variable $Y$ de densité $\pi$ et par conséquent pour presque toute valeur initiale $X_0$ :

$$\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} g(X_t) = \mathbb{E}_Y(g(Y))$$

# Introduction aux méthodes MCMC

- Les MCMC, Méthodes de Monte-Carlo par Chaînes de Markov permettent de simuler numériquement un grand nombre de distributions pour lequelles la densité de probabilité est connue à une constante près.
- Les cas où la loi d'échantillonnage n'est connue qu'au facteur de normalisation près :
  - Physique statistique: la densité de probabilité de trouver le systême dans l'état $\mathbf{x}$ d'énergie $E(\mathbf{x})$ décrit par la distribution de Boltzmann: température $T$ et $k_B$ la constante de Boltzmann :

$$f(\mathbf{x}) \propto \exp\left[-E(\mathbf{x})/k_B T\right]$$

    où le facteur de normalisation est la fonction de partition $Z = \int \exp\left[-E(\mathbf{x})/k_B T\right] d\mathbf{x}$
  - Inférence Bayésienne : loi *a posteriori*

$$\underbrace{\pi(\mathbf{x}|\texttt{observations})}_{\texttt{loi } a \textit{ posteriori}} \propto \underbrace{f(\texttt{observations}|\mathbf{x})}_{\texttt{vraisemblance}} \times \underbrace{\pi(\mathbf{x})}_{\texttt{loi } a \textit{ priori}}$$

- Les MCMC permettent donc d'échantillonner une loi afin d'estimer des grandeurs d'intérêt comme l'espérance, la variance ou un taux d'évènements

# Condition de réversibilité

- On rappelle le cas discret de l'équation de Chapman-Kolmogorov :

$$\pi^{t+1} = \pi^t \boldsymbol{P} \Rightarrow \pi_j^{t+1} = \sum_i \pi_i^t p_{ij}$$

- La passage au cas continu multidimensionnel s'obtient en remplaçant la somme discrète par une intégrale et en notant $p(\mathbf{x} \to \mathbf{y})$ la probabilité de transition de l'état $\mathbf{x}$ à l'état $\mathbf{y}$ :

$$\pi^{t+1}(\mathbf{y}) \quad = \quad \int \pi^t(\mathbf{x}) p(\mathbf{x} \to \mathbf{y}) d\mathbf{x}$$

- On admet les résultats obtenus dans le cas discret : ergodicité $\to$ la distribution invariante est unique (celle qui correspond à la convergence au cours du temps des distributions initiales)
- On note $f(\mathbf{x})$ la densité de probabilité connue à un facteur près
- Théorème : si la loi de transition $p(\mathbf{x} \to \mathbf{y})$ est ergodique et si elle satisfait la condition de réversibilité :

$$f(\mathbf{x}) p(\mathbf{x} \to \mathbf{y}) - f(\mathbf{y}) p(\mathbf{y} \to \mathbf{x}) = 0$$

... alors la distribution de la chaîne converge vers une distribution proportionnelle à $f(\mathbf{x})$

# Condition de réversibilité - Démonstration

- La chaîne étant supposée ergodique, elle converge vers une distribution invariante unique.
- Il suffit de montrer que la distribution $f(\mathbf{x})$ (à un coefficient $\alpha$ près) est invariante :

$$\pi^t(\mathbf{x}) = \alpha f(\mathbf{x}) \Rightarrow \pi^{t+1}(\mathbf{x}) = \alpha f(\mathbf{x})$$

- Démonstration :

$$\begin{aligned}
\pi^{t+1}(\mathbf{x}) &= \int \pi^t(\mathbf{y}) p(\mathbf{y} \to \mathbf{x}) d\mathbf{y} \\
&= \int \alpha f(\mathbf{y}) p(\mathbf{y} \to \mathbf{x}) d\mathbf{y} \\
&= \int \alpha f(\mathbf{x}) p(\mathbf{x} \to \mathbf{y}) d\mathbf{y} \quad \texttt{condition de réversibilité} \\
&= \alpha f(\mathbf{x}) \underbrace{\int p(\mathbf{x} \to \mathbf{y}) d\mathbf{y}}_{=1} \\
&= \alpha f(\mathbf{x})
\end{aligned}$$

- Les MCMC permettent donc de générer une distribution de densité proportionnelle à $f(\mathbf{x})$ à partir d'une probabilité de transition $p(\mathbf{x} \to \mathbf{y})$ vérifiant l'équation

$$f(\mathbf{x}) p(\mathbf{x} \to \mathbf{y}) = f(\mathbf{y}) p(\mathbf{y} \to \mathbf{x})$$

# Algorithme de Metropolis-Hastings

- Objectif : échantillonner selon une loi de proba $f(\mathbf{x})$ que l'on connait à une constante mutiplicative près
- L'algorithme nécessite une valeur initiale $X_0$ et une loi de transition $q(\mathbf{x} \to \mathbf{y})$ appellé aussi loi de proposition *proposal*.

**Algorithm 1** Algorithme de Metropolis-Hastings

**Require:** Condition initiale $X_0$, loi instrumentale $q(\mathbf{x} \to \mathbf{y})$. On pose $t = 0$

1: $\mathbf{x} = X_t$, tirage aléatoire $\mathbf{y}$ avec la loi $q(\mathbf{x} \to \mathbf{y})$
2: $\alpha(\mathbf{x}, \mathbf{y}) = \min[1, \frac{f(\mathbf{y})q(\mathbf{y}\to\mathbf{x})}{f(\mathbf{x})q(\mathbf{x}\to\mathbf{y})}]$
3: tirage $U$ variable uniforme $[0, 1]$
4: **if** $U < \alpha(\mathbf{x}, \mathbf{y})$ **then**
5: $\quad X_{t+1} = \mathbf{y}$ (toujours vrai si $\alpha(\mathbf{x}, \mathbf{y}) = 1$)
6: **else**
7: $\quad X_{t+1} = \mathbf{x}$
8: **end if**
9: $t = t + 1$ retour en 1

- Comme l'algorithme ne depend que du rapport $f(\mathbf{x})/f(\mathbf{y})$, la densité de probabilité $f(\mathbf{x})$ peut donc être connue à une constante près (i.e. le facteur de normalisation ou fonction de partition).

# Démonstration de la condition de réversibilité

- Pour démontrer que $f(\mathbf{x})$ est proportionnelle à la distribution stationnaire de la chaîne de Markov, il suffit de vérifier que la chaîne est réversible par rapport à $f$ :

$$f(\mathbf{x})p(\mathbf{x} \to \mathbf{y}) = f(\mathbf{y})p(\mathbf{y} \to \mathbf{x})$$

- Compte tenu que le tirage $q(\mathbf{x} \to \mathbf{y})$ est accepté avec une probabilité $\alpha(\mathbf{x}, \mathbf{y})$ (paramètre d'une loi Bernoulli), la probabilité de transition est donc :

$$p(\mathbf{x} \to \mathbf{y}) = q(\mathbf{x} \to \mathbf{y})\alpha(\mathbf{x}, \mathbf{y})$$

- Cas $\alpha(\mathbf{x}, \mathbf{y}) = 1 \Rightarrow \alpha(\mathbf{y}, \mathbf{x}) \leq 1$

$$\begin{aligned} f(\mathbf{x})p(\mathbf{x} \to \mathbf{y}) &= f(\mathbf{x})q(\mathbf{x} \to \mathbf{y})\alpha(\mathbf{x}, \mathbf{y}) = f(\mathbf{x})q(\mathbf{x} \to \mathbf{y}) \\ f(\mathbf{y})p(\mathbf{y} \to \mathbf{x}) &= f(\mathbf{y})q(\mathbf{y} \to \mathbf{x})\alpha(\mathbf{y}, \mathbf{x}) = f(\mathbf{y})q(\mathbf{y} \to \mathbf{x})\frac{f(\mathbf{x})q(\mathbf{x} \to \mathbf{y})}{f(\mathbf{y})q(\mathbf{y} \to \mathbf{x})} = f(\mathbf{x})q(\mathbf{x} \to \mathbf{y}) \quad \texttt{(CQFD)} \end{aligned}$$

- Cas $\alpha(\mathbf{x}, \mathbf{y}) < 1 \Rightarrow \alpha(\mathbf{y}, \mathbf{x}) = 1$

$$\begin{aligned} f(\mathbf{x})p(\mathbf{x} \to \mathbf{y}) &= f(\mathbf{x})q(\mathbf{x} \to \mathbf{y})\alpha(\mathbf{x}, \mathbf{y}) = f(\mathbf{x})q(\mathbf{x} \to \mathbf{y})\frac{f(\mathbf{y})q(\mathbf{y} \to \mathbf{x})}{f(\mathbf{x})q(\mathbf{x} \to \mathbf{y})} = f(\mathbf{y})q(\mathbf{y} \to \mathbf{x}) \\ f(\mathbf{y})p(\mathbf{y} \to \mathbf{x}) &= f(\mathbf{y})q(\mathbf{y} \to \mathbf{x})\alpha(\mathbf{y}, \mathbf{x}) = f(\mathbf{y})q(\mathbf{y} \to \mathbf{x}) \quad \texttt{(CQFD)} \end{aligned}$$

# Algorithme de Metropolis-Hastings - Loi de proposition

- Choix de la loi de proposition $q(\mathbf{x}, \mathbf{y})$ influence la qualité de l'algorithme. Le choix se fera pour obtenir (dans la mesure du possible) une exploration rapide de l'espace des états et une convergence vers la distribution stationnaire
- Version de l'algorithme de Metropolis original = la loi instrumentale est symétrique

$$q(\mathbf{x} \to \mathbf{y}) = q(\mathbf{y} \to \mathbf{x})$$

- Le rapport des probabilités se simplifie :

$$\alpha(\mathbf{x}, \mathbf{y}) = \min[1, \frac{f(\mathbf{y})}{f(\mathbf{x})}]$$

- Version de l'algorithme de Metropolis-Hastings indépendant : la loi de transition ne dépend pas de l'état courant :

$$q(\mathbf{x} \to \mathbf{y}) = q(\mathbf{y}) \Rightarrow \alpha(\mathbf{x}, \mathbf{y}) = \min[1, \frac{f(\mathbf{y})q(\mathbf{x})}{f(\mathbf{x})q(\mathbf{y})}]$$

- Remarques : dans ce domaine un large champ de recherche notamment pour ajuster/adapter la loi instrumentale $q(\mathbf{x} \to \mathbf{y})$ au cours des tirages Monte Carlo

# Metropolis-Hastings, pour ou contre ?

## Avantages:

- Très simple & très général
- Permet l'échantillonnage selon une grande variété de distributions de probabilité

# Metropolis-Hastings, pour ou contre ?

## Avantages:

- Très simple & très général
- Permet l'échantillonnage selon une grande variété de distributions de probabilité

## Inconvénients:

- Le choix du *proposal* est crucial, c'est le degré de liberté principal de l'algorithme
- Fléau de la dimension
- Seulement des heuristiques pour vérifier la convergence de la chaîne de Markov vers sa distribution stationnaire

# Convergence des MCMC

L'objectif du MCMC est d'échantillonner selon $f$ connue à une constante multiplicative près

# Convergence des MCMC

L'objectif du MCMC est d'échantillonner selon $f$ connue à une constante multiplicative près

Aucune garantie de la convergence de la chaîne en temps fini !

# Convergence des MCMC

L'objectif du MCMC est d'échantillonner selon $f$ connue à une constante multiplicative près

Aucune garantie de la convergence de la chaîne en temps fini !

Il existe de nombreuses "astuces" pour à la fois s'assurer de la convergence de la chaîne et de l'accélérer:

- *Burn-in*
- *Thinning*
- Autocorrélation
- Taille d'échantillon effective (*Effective Sample size* ou ESS)

# Application: segmentation TEP

La loi *a priori* pour chaque zone d'une image TEP est le champ de Potts:

$$\pi(\mathbf{z}) \propto \exp\left[\sum_{i=1}^{n} \sum_{i' \in \mathcal{V}(i)} \gamma \mathbf{1}_{z_i = z_{i'}}\right]$$

On se place dans le cas où il n'existe que 2 zones $z_i \in \{+1, -1\}$.

## Application: segmentation TEP

La loi *a priori* pour chaque zone d'une image TEP est le champ de Potts:

$$\pi(\mathbf{z}) \propto \exp\left[\sum_{i=1}^{n} \sum_{i' \in \mathcal{V}(i)} \gamma \mathbf{1}_{z_i = z_{i'}}\right]$$

On se place dans le cas où il n'existe que 2 zones $z_i \in \{+1, -1\}$.

Dans ce cas, le champ de Potts est équivalent au modèle d'Ising, très utilisé en physique statistique.

# Application: segmentation TEP

La loi *a priori* pour chaque zone d'une image TEP est le champ de Potts:

$$\pi(\mathbf{z}) \propto \exp\left[\sum_{i=1}^{n} \sum_{i' \in \mathcal{V}(i)} \gamma \mathbf{1}_{z_i = z_{i'}}\right]$$

On se place dans le cas où il n'existe que 2 zones $z_i \in \{+1, -1\}$.

Dans ce cas, le champ de Potts est équivalent au modèle d'Ising, très utilisé en physique statistique.

Pour une image de taille $n$, la constante de normalisation est

$$C(\gamma) = \sum_{(z_1,\ldots,z_n) \in \{+1,-1\}^n} \exp\left[\sum_{i=1}^{n} \sum_{i' \in \mathcal{V}(i)} \gamma \mathbf{1}_{z_i = z_{i'}}\right]$$

Il faut sommer $2^n$ composantes ! Très coûteux pour des images TEP grandes !

$\hookrightarrow$ Nécessité d'utiliser des techniques MCMC pour l'échantillonnage

# Algorithme Metropolis-Hastings pour le champ de Potts

On remarque que $\pi(\mathbf{z}) \propto \exp[U(\gamma, \mathbf{z})]$ avec $U$ une "fonction d'utilité".

# Algorithme Metropolis-Hastings pour le champ de Potts

On remarque que $\pi(\mathbf{z}) \propto \exp[U(\gamma, \mathbf{z})]$ avec $U$ une "fonction d'utilité".

On utilise comme *proposal* la procédure suivante:

- on choisit un pixel au hasard
- on change sa catégorie (i.e. si $z_i = +1$ alors il devient $-1$)

# Algorithme Metropolis-Hastings pour le champ de Potts

On remarque que $\pi(\mathbf{z}) \propto \exp[U(\gamma, \mathbf{z})]$ avec $U$ une "fonction d'utilité".

On utilise comme *proposal* la procédure suivante:

- on choisit un pixel au hasard
- on change sa catégorie (i.e. si $z_i = +1$ alors il devient $-1$)

On calcule entre l'état précédent et l'état suivant le ratio d'acceptation:

$$\frac{\exp[U_{\text{next}}]}{\exp[U]} = \exp[\Delta U]$$

On peut calculer que

$$\Delta U = -\gamma * z_i * \sum_{i' \in \mathcal{V}_i} z_{i'}$$

Le nouvel état est accepté si $\Delta U > 0$ ou bien avec probabilité $\exp[\Delta U]$.

# Illustration

# Références


- The beginning of the Monte Carlo method, N. Metropolis, *Los Alamos Science* special issue, 1987
- Exemple de simulation MCMC
  `https://chi-feng.github.io/mcmc-demo/app.html?algorithm=RandomWalkMH&target=banana`