

MAPSI — cours 10 : Échantillonnage et MCMC

Pierre-Henri Wuillemin & Christophe Gonzales

LIP6 / ISIR – Sorbonne Université, France

Échantillonnage en statistique

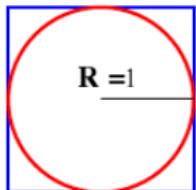
L'échantillonnage (méthodes de Monte Carlo) proposent une **simulation stochastique** pour le calcul d'intégrales (ou d'équations différentielles).

$$E_P(f) = \int f(x) \cdot P(x) dx \approx \frac{1}{N} \sum_{i \leq N} f(X_i), \forall i, X_i \text{ iid, suivant } P$$

Il s'avère que l'intégration (ou l'équivalent discret : **la somme**) est une opération fondamentale dans les statistiques :

- À partir de $posterior \propto L \times P$, calculer la constante de normalisation (Z) : $\int L \times P$ car $posterior = \frac{L \times P}{\int L \times P}$
- Marginaliser une distribution jointe : $P(x_2) = \int P(x_1, x_2) dx_1$
- Statistiques sur une distribution :
 - Moyenne de P : $f(x) = x$
 - Moment d'ordre 2 de P : $f(x) = x^2$
 - $P(A)$: $f(X) = \mathbf{1}_A$

Calcul (un peu tiré par les cheveux) de π par échantillonnage



$$\pi = \int_{\bigcirc} dx = \int_{\square} 1_{\bigcirc}(x) dx \propto \int_{\square} p(x).1_{\bigcirc}(x) dx$$

où p distribution uniforme sur \square .

Méthode : on jette des cailloux dans le carré.

Hypothèse : les jets suivent une distribution uniformes
d'où

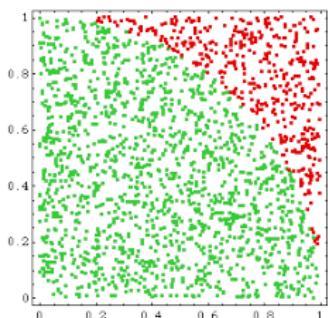
$$\frac{1}{N} \sum_{i \leq N} 1_{\bigcirc}(X_i) = \frac{\text{NbJets}_{\text{Cercle}}}{\text{NbJets}_{\text{Total}}} = \frac{\pi \cdot R^2}{(2 \cdot R)^2} = \frac{\pi}{4}$$

6 jets dans le cercle sur 10 jets en tout $\Rightarrow \hat{\pi}_{10} = 2.4$

89 jets dans le cercle sur 100 jets en tout $\Rightarrow \hat{\pi}_{100} = 3.57$

750 jets dans le cercle sur 1000 jets en tout $\Rightarrow \hat{\pi}_{1000} = 3$

7852 jets dans le cercle sur 10000 jets en tout $\Rightarrow \hat{\pi}_{10000} = 3.1408$



- 1 Échantillonnage d'une loi discrète
- 2 Rejection sampling
- 3 MCMC : Metropolis-Hastings
- 4 MCMC : échantillonneur de Gibbs

Échantillonnage avec une distribution discrète

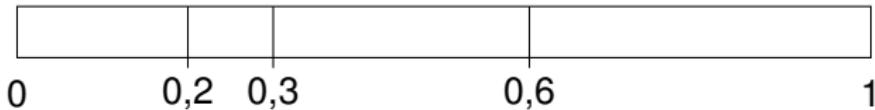
- *Problème :* échantillonner selon :

	X_1	X_2	X_3	X_4
distribution $\pi^\infty(X)$ =	0,2	0,1	0,3	0,4

- *Solution :*

- 1 Calculer la cumulative :

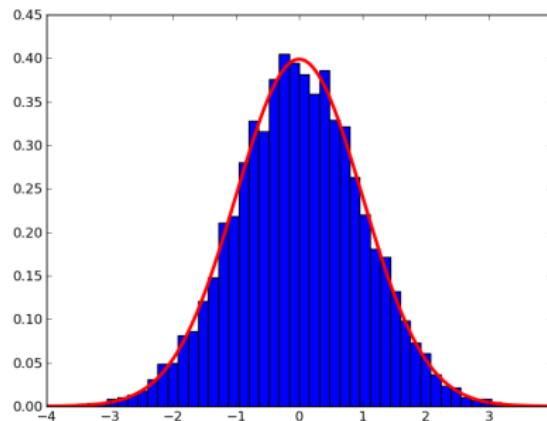
$$F(X_i) = \sum_{Y \leq X_i} \pi^\infty(Y) = \boxed{0,2 \quad 0,3 \quad 0,6 \quad 1}$$



- 2 Tirer un nombre z selon une distribution uniforme sur $[0,1[$
- 3 Soit i tel que $F(X_{i-1}) \leq z < F(X_i)$ (en posant $X_0 = 0$)
- 4 Renvoyer X_i

Échantillonnage d'une loi normale

- Faire la cumulative de la fonction de densité (cf. table)



Il existe des algos dédiés performants
(Ziggurat, Box-Muller)

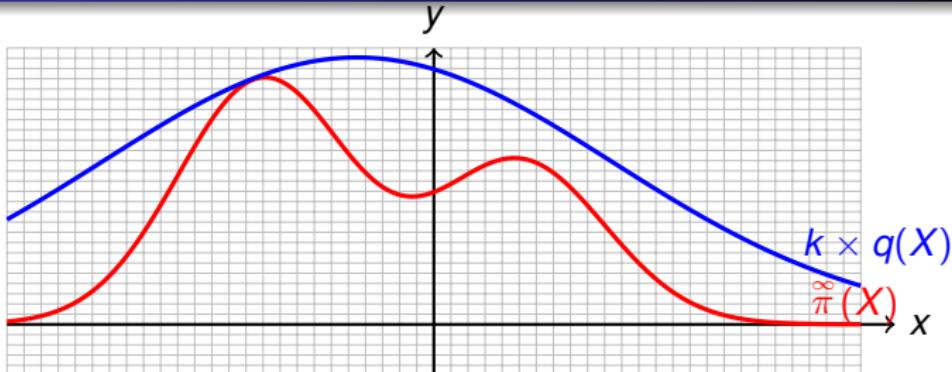
Distributions complexes : méthode d'inversion

- Variable aléatoire continue x de densité f et fonction de répartition F
- $F(x) = P(X \leq x) = \int_{y=-\infty}^x f(y)dy$
⇒ Inverser la fonction de répartition : $x = F^{-1}(u)$

Échantillonnage de $x \sim f$

- ➊ Tirer $u \sim U([0, 1])$ - loi uniforme sur $[0, 1]$
 - ➋ Appliquer $x = F^{-1}(u)$
⇒ x distribué selon f (preuve en TD)
- Nécessite la forme analytique de l'inverse F^{-1}

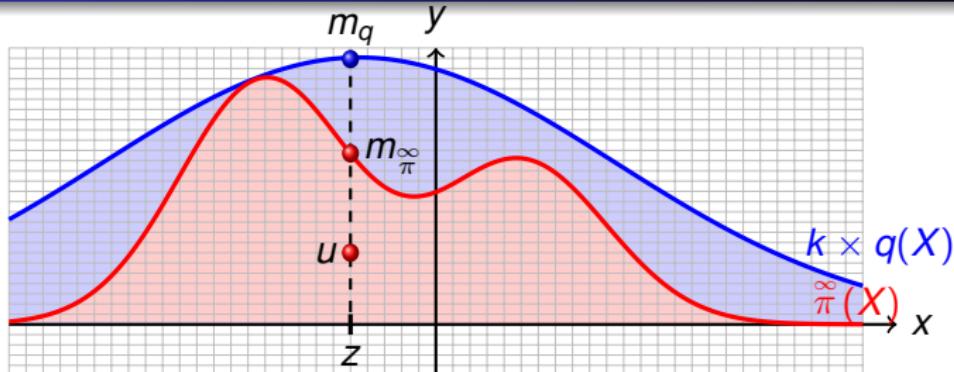
Distributions complexes : Rejection Sampling



Hypothèses :

- $\pi^\infty(\cdot)$ difficile à échantillonner
- Mais pour tout $x \in X$, $\pi^\infty(x)$ facile à calculer
- $q(\cdot)$ facile à échantillonner
- il existe $k \in \mathbb{R}$ tel que $\pi^\infty(x) \leq k \times q(x)$ pour tout $x \in X$

Distributions complexes : Rejection Sampling



Algorithme « rejection sampling » :

- ➊ Tirer un nombre z selon $q(\cdot)$
- ➋ Calculer $m_q = k \times q(z)$
- ➌ Tirer un nombre u selon une loi uniforme sur $[0, m_q[$
- ➍ Accepter z si $u \leq \pi(z) = m_{\pi}^{\infty}$

Rejection Sampling : preuve

- $P(X \leq x | X \text{ accepté}) = \frac{P(X \leq x, X \text{ accepté})}{P(X \text{ accepté})}$
- Calcul du taux d'acceptation :
$$\begin{aligned} P(\text{acceptation}) &= \int_{-\infty}^{+\infty} q(z) \times \frac{m_{\pi}^{\infty}(z)}{m_q(z)} dz \\ &= \int_{-\infty}^{+\infty} q(z) \times \frac{\pi^{\infty}(z)}{k \times q(z)} dz \\ &= \frac{1}{k} \int_{-\infty}^{+\infty} \pi^{\infty}(z) dz = \frac{1}{k} \end{aligned}$$
- $P(X \leq x, X \text{ accepté}) = \int_{-\infty}^x \frac{\pi^{\infty}(z)}{k \times q(z)} q(z) dz = \frac{F(x)}{k}$

Rejection Sampling

Avantage : fonction de partition inconnue

- $\hat{\pi}(x) = \frac{1}{Z_p} p(x)$
 - Seul $p(x)$ connu
 - **Nouvelle règle :** $k \times q(x) \geq p(x)$ pour tout x
 - Rejection sampling \Rightarrow échantillon $\langle z_1, \dots, z_n \rangle \sim \hat{\pi}(\cdot)$
 - $\hat{\pi}(z) \propto q(z) \times \frac{p(z)}{k \times q(z)}$
 - $\hat{\pi}(z) \propto \frac{p(z)}{k} \propto p(z) \propto \hat{\pi}(z)$
- \Rightarrow on peut échantillonner sans connaître la fonction de partition

Calcul du taux d'acceptation :

$$\begin{aligned} P(\text{acceptation}) &= \int q(z) \times \frac{m_{\pi}^{\infty}(z)}{m_q(z)} dz \\ &= \int q(z) \times \frac{\pi^{\infty}(z)}{k \times q(z)} dz \\ &= \frac{1}{k} \int \pi^{\infty}(z) dz = \frac{1}{k} \end{aligned}$$



Exemple précédent : $k \approx 2 \implies$ seulement 1 z sur 2 accepté !



k augmente exponentiellement avec la dimension de $\pi^{\infty}(\cdot)$!

Principe de l'échantillonnage par MCMC

MCMC : Markov Chain Monte Carlo

- **But** : échantillonner selon une loi $\tilde{\pi}(\cdot)$
- **Principe** : construire une suite (X_i) de variables aléatoires tirées selon des lois $\tilde{\pi}_i(\cdot)$ tendant vers $\tilde{\pi}(\cdot)$
et sélectionner un échantillon $\langle x_i, \dots, x_{m+i} \rangle$
ou sous-échantillonner : $\langle x_{\sigma(i)}, \dots, x_{\sigma(m+i)} \rangle \implies \approx \text{i.i.d.}$
- **Solution** : construire une chaîne de Markov de loi stationnaire $\tilde{\pi}(\cdot)$

- soit $P(X_{t+1}|X_t)$ la probabilité de transition (chaîne homogène)

Loi stationnaire $\pi^\infty(\cdot)$

$$\pi^\infty(x) = \int_y P(x|y) \pi^\infty(y) dy$$



ici, on connaît $\pi^\infty(\cdot)$ et on cherche $P(\cdot|\cdot)$

Problème : sous quelles conditions $P(\cdot|\cdot)$ existe-t-elle ?

Ergodicité ?

Une condition possible

Réversibilité

$$\pi^\infty(x)P(y|x) = \pi^\infty(y)P(x|y), \forall x, y$$



propriété également connue sous le nom de
« detailed balance »

conséquence :

$$\begin{aligned} \int_y P(x|y) \pi^\infty(y) dy &= \int_y P(y|x) \pi^\infty(x) dy \\ &= \pi^\infty(x) \int_y P(y|x) dy \\ &= \pi^\infty(x) \end{aligned}$$

$\implies \pi^\infty(\cdot)$ loi stationnaire !

Garantir la réversibilité

- En général, $\tilde{\pi}^\infty(x)P(y|x) \neq \tilde{\pi}^\infty(y)P(x|y)$

Interprétation de $\tilde{\pi}^\infty(x)P(y|x) > \tilde{\pi}^\infty(y)P(x|y)$

Le processus markovien va évoluer plus souvent de x vers y que de y vers $x \implies$ non réversible.

Correction : diminuer $P(y|x)$ ou augmenter $P(x|y)$

\implies créer deux nombres $\alpha(x,y)$ et $\alpha(y,x)$ tels que :

$$\tilde{\pi}^\infty(x)P(y|x)\alpha(x,y) = \tilde{\pi}^\infty(y)P(x|y)\alpha(y,x)$$



on veut que $P(y|x)\alpha(x,y)$ soit une probabilité de transition !

Remarque : $y = x \implies \tilde{\pi}^\infty(x)P(x|x)\alpha(x,x) = \tilde{\pi}^\infty(x)P(x|x)\alpha(x,x)$
pour tout $\alpha(x,x)$

Si $P(x|x)\alpha(x,x) = 1 - \int_{y \neq x} P(y|x)\alpha(x,y)dy$, on a bien une proba !

Garantir la réversibilité

$$P(x|x)\alpha(x,x) = 1 - \int_{y \neq x} P(y|x)\alpha(x,y)dy$$

Pour assurer que $P(x|x)\alpha(x,x) \geq 0$, on impose $\alpha(x,y) \leq 1$

Supposons $\bar{\pi}(x)P(y|x) > \bar{\pi}(y)P(x|y)$
puisque $\bar{\pi}(x)P(y|x)\alpha(x,y) = \bar{\pi}(y)P(x|y)\alpha(y,x)$

Il faut "augmenter" $P(x|y)$.

\implies pour augmenter $P(x|y)$, on fixe $\alpha(y,x) = 1$

$\implies \bar{\pi}(x)P(y|x)\alpha(x,y) = \bar{\pi}(y)P(x|y)$

$$\implies \alpha(x,y) = \frac{\bar{\pi}(y)P(x|y)}{\bar{\pi}(x)P(y|x)}$$

Résumé :

- Si $\hat{\pi}(x)P(y|x) > \hat{\pi}(y)P(x|y)$:

Fixer $\alpha(x,y) = \frac{\hat{\pi}(y)P(x|y)}{\hat{\pi}(x)P(y|x)}$ et $\alpha(y,x) = 1$

- Par symétrie, si $\hat{\pi}(x)P(y|x) < \hat{\pi}(y)P(x|y)$:

Fixer $\alpha(x,y) = 1$ et $\alpha(y,x) = \frac{\hat{\pi}(x)P(y|x)}{\hat{\pi}(y)P(x|y)}$

Interprétation de α : probabilité de mouvement

$\alpha(x, y) =$ la probabilité de **réaliser** la transition de x vers y

\implies à l'étape t , on a 2 choix :

- transiter de x vers un y avec la probabilité $P(y|x)\alpha(x, y)$
- ne pas réaliser de transition

Résumé

si $\alpha(x, y) = \min \left\{ 1, \frac{\pi^\infty(y)P(x|y)}{\pi^\infty(x)P(y|x)} \right\}$ alors :

$$\pi^\infty(x)P(y|x)\alpha(x, y) = \pi^\infty(y)P(x|y)\alpha(y, x)$$

\implies réversibilité $\implies \pi^\infty(\cdot)$ distribution stationnaire

Algorithme de Metropolis-Hastings

Metropolis-Hastings

Algorithme pour générer x_{t+1} à partir de x_t :

- ➊ tirer z selon la distribution $P(\cdot|x_t)$
- ➋ calculer $\alpha(x_t, z) = \min \left\{ 1, \frac{\pi(z)P(x_t|z)}{\pi(x_t)P(z|x_t)} \right\}$
- ➌ tirer un nombre u selon une loi uniforme sur $[0, 1[$
- ➍ renvoyer $x_{t+1} = \begin{cases} z & \text{si } u \leq \alpha(x_t, z) \\ x_t & \text{sinon} \end{cases}$

Références :

- ➊ N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller et E. Teller (1953) “Equations of State Calculations by Fast Computing Machines”. Journal of Chemical Physics, 21 (6), pp. 1087–1092
- ➋ W.K. Hastings (1970) “Monte Carlo Sampling Methods Using Markov Chains and Their Applications”. Biometrika, 57 (1), pp. 97–109

- $P(\cdot|x_t)$ doit être simple à échantillonner
- 1ère possibilité [Metropolis *et al.* (1953), Müller (1993)]

$P(z|x_t) = q(z - x_t)$ avec $q(\cdot)$ densité multivariée
autrement dit $z = x_t + y$ avec $y \sim q(\cdot)$

 $q(\cdot)$ indépendante de x_t !

⇒ random walk chain

- choix possible de $q(\cdot)$: loi normale
- si q est symétrique : $q(y) = q(-y)$ et

$$\alpha(x_t, z) = \min \left\{ 1, \frac{\pi^\infty(z)P(x_t|z)}{\pi^\infty(x_t)P(z|x_t)} \right\} = \min \left\{ 1, \frac{\pi^\infty(z)}{\pi^\infty(x_t)} \right\}$$

- 2ème possibilité [Hastings (1970)]

$P(z|x) = q(z)$ avec $q(\cdot)$ densité multivariée

⇒ independent chain

⇒ généralisation de rejection sampling

- 3ème possibilité : l'algorithme Langevin [Roberts et Rosenthal (1998)]

$$z = x_t + \frac{\sigma^2}{2} \nabla \log(\tilde{\pi}(x_t)) + \sigma y \text{ avec } y \sim q(\cdot)$$

σ : facteur d'échelle



Il existe plein d'autres possibilités...



important pour la vitesse de convergence

Influence de l'étalement

- Taux d'acceptation
- Région couverte par la chaîne de Markov

Choix de l'étalement/variance de $P(\cdot|x_t)$

Roberts, Gelman, Gilks (1994)

- cadre : random walk
- π^∞ et $P(\cdot|\cdot)$: lois normales mono-dimensionnelles
 - affiner l'étalement de $P(\cdot|x_t)$ pour obtenir un taux d'acceptation $\approx 0,45$
- π^∞ et $P(\cdot|\cdot)$: lois normales n -dimensionnelles
 - affiner l'étalement de $P(\cdot|x_t)$ pour obtenir un taux d'acceptation $\approx 0,23$ lorsque n tend vers $+\infty$

Müller (1993)

Random walk \implies taux d'acceptation $\approx 0,5$.

Initialisation et burn in

Initialisation :

Partir de n'importe quelle valeur x_0



au départ l'échantillon ne suit pas $\pi^\infty(\cdot)$

⇒ burn in nécessaire :

Ne conserver dans l'échantillon que les x_t pour $t > t_0$

En général, t_0 est de l'ordre de quelques milliers

- supposons que $x_t = (x_t^1, x_t^2)$

Précédemment :

- stationnarité : $\tilde{\pi}(x_{t+1}) = \int_{x_t} P(x_{t+1}|x_t) \tilde{\pi}(x_t) dx_t$

Maintenant :

- $P(x_{t+1}|x_t) = P(x_{t+1}^1, x_{t+1}^2|x_t^1, x_t^2)$

- stationnarité :

$$\tilde{\pi}(x_{t+1}^1, x_{t+1}^2) = \int_{x_t^1} \int_{x_t^2} P(x_{t+1}^1, x_{t+1}^2|x_t^1, x_t^2) \tilde{\pi}(x_t^1, x_t^2) dx_t^1 dx_t^2$$

- Or $P(x_{t+1}|x_t) = P(x_{t+1}^2|x_{t+1}^1, x_t^1, x_t^2) \times P(x_{t+1}^1|x_t^1, x_t^2)$
 $= P(x_{t+1}^2|x_{t+1}^1, x_t^2) \times P(x_{t+1}^1|x_t^1, x_t^2)$ (prop. Markov)

Metropolis-Hastings par bloc : stationnarité

$$\pi^\infty(x_{t+1}^2, x_{t+1}^2) = \int_{x_t^1} \int_{x_t^2} P(x_{t+1}^2 | x_{t+1}^1, x_t^2) \times P(x_{t+1}^1 | x_t^1, x_t^2) \pi^\infty(x_t^1, x_t^2) dx_t^1 dx_t^2$$

Rappel : stationnarité pour 1 variable

$$\pi^\infty(x_{t+1}) = \int_{x_t} P(x_{t+1} | x_t) \pi^\infty(x_t) dx_t$$

Stationnarité par bloc

Généralisation en rajoutant toutes les variables sauf celle en x_{t+1}^i à droite des signes de conditionnement :

- $\pi^\infty(x_{t+1}^1 | y^2) = \int_{x_t^1} P(x_{t+1}^1 | x_t^1, y^2) \pi^\infty(x_t^1 | y^2) dx_t^1$ pour tout y^2
- $\pi^\infty(x_{t+1}^2 | y^1) = \int_{x_t^2} P(x_{t+1}^2 | x_t^2, y^1) \pi^\infty(x_t^2 | y^1) dx_t^2$ pour tout y^1

Metropolis-Hastings par bloc : stationnarité

conséquences de la stationnarité par bloc :

$$\int_{x_t^1} \int_{x_t^2} P(x_{t+1}^2 | x_{t+1}^1, x_t^2) \times P(x_{t+1}^1 | x_t^1, x_t^2) \stackrel{\infty}{\pi}(x_t^1, x_t^2) dx_t^1 dx_t^2$$

$$= \int_{x_t^1} \int_{x_t^2} P(x_{t+1}^2 | x_{t+1}^1, x_t^2) \times P(x_{t+1}^1 | x_t^1, x_t^2) \stackrel{\infty}{\pi}(x_t^1 | x_t^2) \stackrel{\infty}{\pi}(x_t^2) dx_t^1 dx_t^2$$

$$= \int_{x_t^2} \int_{x_t^1} P(x_{t+1}^2 | x_{t+1}^1, x_t^2) \times P(x_{t+1}^1 | x_t^1, x_t^2) \stackrel{\infty}{\pi}(x_t^1 | x_t^2) \stackrel{\infty}{\pi}(x_t^2) dx_t^1 dx_t^2$$

$$= \int_{x_t^2} P(x_{t+1}^2 | x_{t+1}^1, x_t^2) \left[\int_{x_t^1} P(x_{t+1}^1 | x_t^1, x_t^2) \stackrel{\infty}{\pi}(x_t^1 | x_t^2) dx_t^1 \right] \stackrel{\infty}{\pi}(x_t^2) dx_t^2$$

$$= \int_{x_t^2} P(x_{t+1}^2 | x_{t+1}^1, x_t^2) \stackrel{\infty}{\pi}(x_{t+1}^1 | x_t^2) \stackrel{\infty}{\pi}(x_t^2) dx_t^2 \quad (\text{stationnarité par bloc})$$

$$= \int_{x_t^2} P(x_{t+1}^2 | x_{t+1}^1, x_t^2) \stackrel{\infty}{\pi}(x_t^2 | x_{t+1}^1) \stackrel{\infty}{\pi}(x_{t+1}^1) dx_t^2 \quad (\text{formule de Bayes})$$

$$= \stackrel{\infty}{\pi}(x_{t+1}^2 | x_{t+1}^1) \stackrel{\infty}{\pi}(x_{t+1}^1) = \stackrel{\infty}{\pi}(x_{t+1}^1, x_{t+1}^2) \quad (\text{stationnarité par bloc})$$

Metropolis-Hastings par bloc

Conclusion du transparent précédent

Stationnarité par bloc \implies Stationnarité de la loi jointe

Metropolis-Hastings par bloc

Algorithme pour générer $x_{t+1} = (x_{t+1}^1, \dots, x_{t+1}^n)$ à partir de x_t :

1 choisir une permutation $\sigma : \{1, \dots, n\} \mapsto \{1, \dots, n\}$

2 pour tout $i \in \{1, \dots, n\}$ faire :

a Posons $y = (x_{t+1}^{\sigma(1)}, \dots, x_{t+1}^{\sigma(i-1)}, x_t^{\sigma(i+1)}, \dots, x_t^{\sigma(n)})$

b tirer $z^{\sigma(i)}$ selon la distribution $P(\cdot | x_t^{\sigma(i)}, y)$

c calculer $\alpha(x_t^{\sigma(i)}, z^{\sigma(i)} | y) = \min \left\{ 1, \frac{\pi(z^{\sigma(i)} | y) P(x_t^{\sigma(i)} | z^{\sigma(i)}, y)}{\pi(x_t^{\sigma(i)} | y) P(z^{\sigma(i)} | x_t^{\sigma(i)}, y)} \right\}$

d tirer un nombre u selon une loi uniforme sur $[0, 1]$

e $x_{t+1}^{\sigma(i)} = \begin{cases} z^{\sigma(i)} & \text{si } u \leq \alpha(x_t^{\sigma(i)}, z^{\sigma(i)} | y) \\ x_t^{\sigma(i)} & \text{sinon} \end{cases}$

Échantillonneur de Gibbs

- Metropolis-Hastings par bloc
- Choix de la proba de transition : $P(z^{\sigma(i)}|x_t^{\sigma(i)}, y) = \pi^\infty(z^{\sigma(i)}|y)$

Conséquence :

$$\alpha(x_t^{\sigma(i)}, z^{\sigma(i)}|y) = \min \left\{ 1, \frac{\pi^\infty(z^{\sigma(i)}|y)P(x_t^{\sigma(i)}|z^{\sigma(i)}, y)}{\pi^\infty(x_t^{\sigma(i)}|y)P(z^{\sigma(i)}|x_t^{\sigma(i)}, y)} \right\} = 1$$

$\implies z^{\sigma(i)}$ est toujours accepté

Algorithme

Algorithme pour générer $x_{t+1} = (x_{t+1}^1, \dots, x_{t+1}^n)$ à partir de x_t :

- ① choisir une permutation $\sigma : \{1, \dots, n\} \mapsto \{1, \dots, n\}$
- ② pour tout $i \in \{1, \dots, n\}$ faire :
 - ⓐ posons $y = (x_{t+1}^{\sigma(1)}, \dots, x_{t+1}^{\sigma(i-1)}, x_t^{\sigma(i+1)}, \dots, x_t^{\sigma(n)})$
 - ⓑ tirer $x_{t+1}^{\sigma(i)}$ selon la distribution $\tilde{\pi}(\cdot | y)$