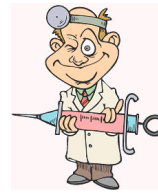


MAPSI — cours 9 : Échantillonnage et MCMC

Christophe Gonzales

LIP6 – Université Paris 6, France

Motivations : monitoring de patients



Réseau « alarm » :

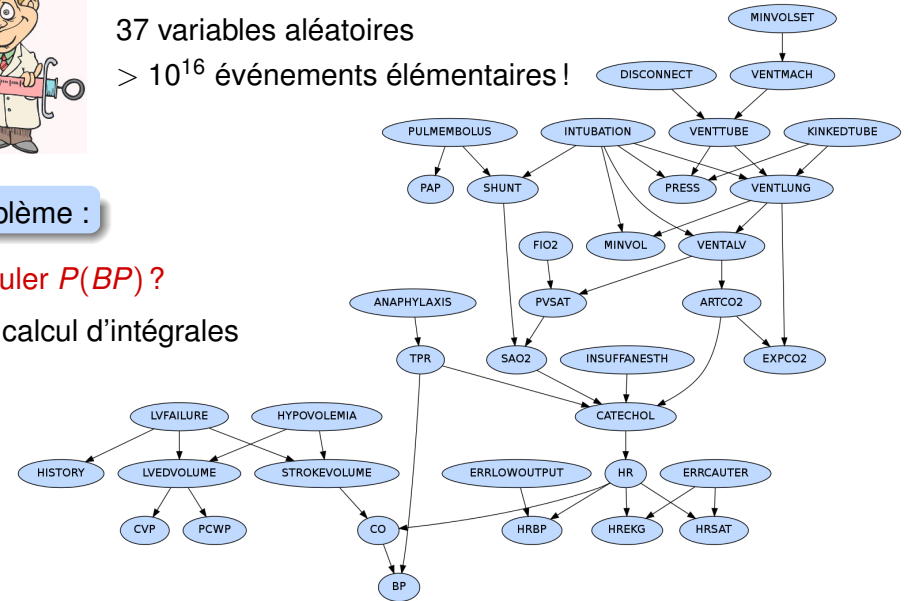
37 variables aléatoires

> 10¹⁶ événements élémentaires !

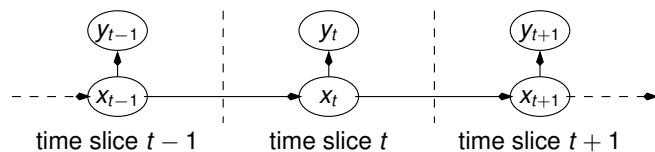
Problème :

calculer $P(BP)$?

⇒ calcul d'intégrales

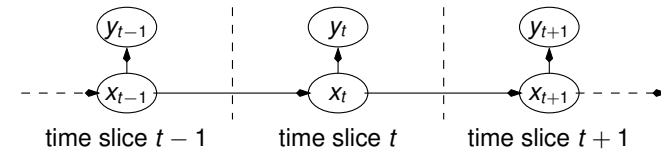


Motivations : vidéo tracking



● **But** : Estimer x_t sachant $y_{1:t}$ pour tout t : $p(x_t|y_{1:t})$

Motivations : vidéo tracking



1 Prediction : $p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})dx_{t-1}$

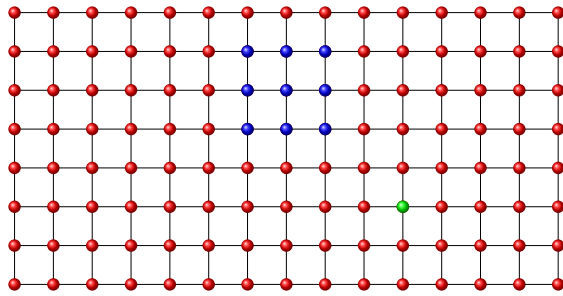
$$P(A|B) = \frac{P(A, B)}{P(B)}$$

$$P(A, C|B) = \frac{P(A, C, B)}{P(B)} = \frac{P(A, C, B)}{P(C, B)} \frac{P(C, B)}{P(B)} = P(A|C, B)P(C|B)$$

$$P(A|B) = \int P(A, C|B)dC$$

2 Correction : $p(x_t|y_{1:t}) \propto p(y_t|x_t)p(x_t|y_{1:t-1})$

Motivations : modèles d'Ising



- **Énergie** : $E = \sum_i \sum_{j \text{ voisin de } i} \psi(i,j) + \sum_i \phi(i)$
- **Probabilité d'une configuration** : $P = \frac{1}{Z} e^{-\beta E}$

Problèmes : calculer Z et calculer E ?

- **Applications** : magnétisme, gaz, neuroscience, nouveaux modèles probabilistes...

Plan du cours n°9

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

Échantillonnage avec une distribution discrète

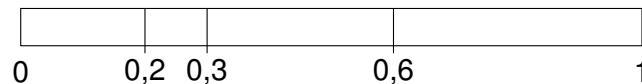
- **Problème** : échantillonner selon :

$$\text{distribution } \pi(X) = \begin{array}{|c|c|c|c|} \hline X_1 & X_2 & X_3 & X_4 \\ \hline 0,2 & 0,1 & 0,3 & 0,4 \\ \hline \end{array}$$

- **Solution** :

- 1 Calculer la cumulative :

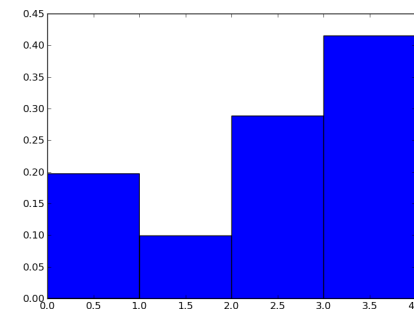
$$F(X_i) = \sum_{Y \leq X_i} \pi(Y) = \begin{array}{|c|c|c|c|} \hline 0,2 & 0,3 & 0,6 & 1 \\ \hline \end{array}$$



- 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

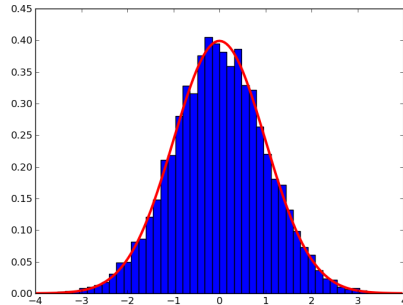
Échantillonnage avec une distribution discrète

$$\pi(X) = \begin{array}{|c|c|c|c|} \hline X_1 & X_2 & X_3 & X_4 \\ \hline 0,2 & 0,1 & 0,3 & 0,4 \\ \hline \end{array}$$

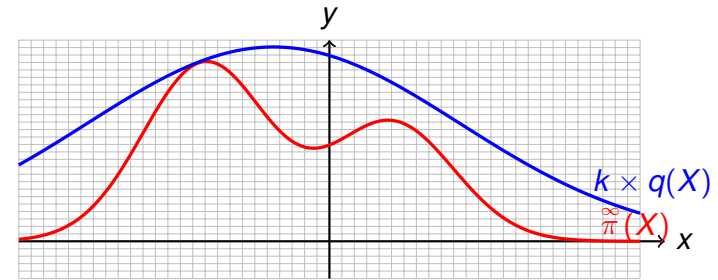


Échantillonnage d'une loi normale

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

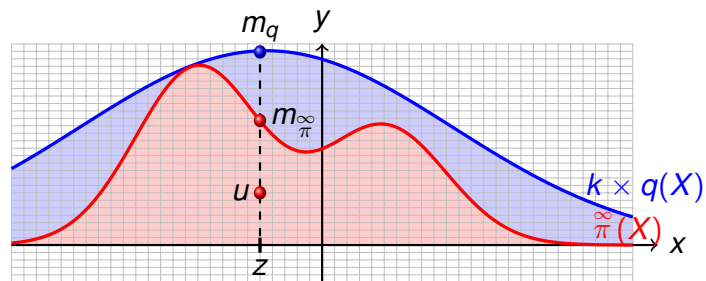


⚠ Il existe des algos dédiés performants (Ziggurat, Box-Muller)



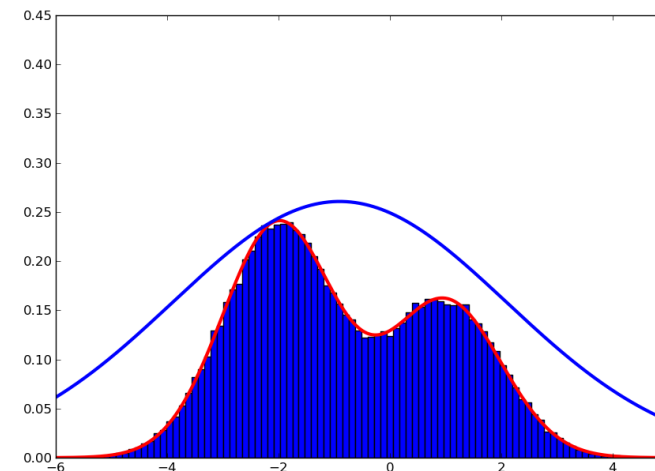
Hypothèses :

- $\pi^\infty(\cdot)$ 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 $x \in X$



Algorithme « rejection sampling » :

- 1 Tirer un nombre z selon $q(\cdot)$
- 2 Calculer $m_q = k \times q(z)$
- 3 Tirer un nombre u selon une loi uniforme sur $[0, m_q[$
- 3 Accepter z si $u \leq \pi^\infty(z) = m_\pi$





Avantage : fonction de partition inconnue

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

Calcul du taux d'acceptation :

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

 Exemple précédent : $k = 1,96 \implies$ seulement 1 z sur 2 accepté !

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

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 : $\langle X_{\sigma(i)}, \dots, X_{\sigma(m+i)} \rangle \implies \approx$ i.i.d.
- **Solution** : construire une chaîne de Markov de loi stationnaire $\tilde{\pi}(\cdot)$

Loi stationnaire

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

Loi stationnaire $\tilde{\pi}(\cdot)$

$$\tilde{\pi}(x) = \int_y P(x|y) \tilde{\pi}(y) dy$$

 ici, on connaît $\tilde{\pi}(\cdot)$ et on cherche $P(\cdot|\cdot)$

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

Une condition possible

Réversibilité

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

⚠ propriété également connue sous le nom de « detailed balance »

conséquence :

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

⇒ $\pi(\cdot)$ loi stationnaire!

Garantir la réversibilité

- En général, $\pi(x)P(y|x) \neq \pi(y)P(x|y)$

Interprétation de $\pi(x)P(y|x) > \pi(y)P(x|y)$

Le processus markovien va évoluer plus souvent de x vers y que de y vers x ⇒ non réversible.

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

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

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

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

Remarque : $y = x \Rightarrow \pi(x)P(x|x)\alpha(x, x) = \pi(x)P(x|x)\alpha(x, x)$
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 $\alpha(x, y) \leq 1$

$$\begin{aligned} \pi(x)P(y|x) &> \pi(y)P(x|y) \\ \pi(x)P(y|x)\alpha(x, y) &= \pi(y)P(x|y)\alpha(y, x) \end{aligned}$$

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

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

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

Garantir la réversibilité

Résumé :

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

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

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

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

Interprétation de α : probabilité de mouvement

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

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

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

⇒ réversibilité ⇒ $\pi(\cdot)$ distribution stationnaire

Metropolis-Hastings

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

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

Références :

- 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
- W.K. Hastings (1970) "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". Biometrika, 57 (1), pp. 97–109

Choix de $P(\cdot|x_t)$

- $P(\cdot|x_t)$ 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

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

Choix de $P(\cdot|x_t)$

- 2ème possibilité [Hastings (1970)]

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


⇒ independent chain

⇒ généralisation de rejection sampling

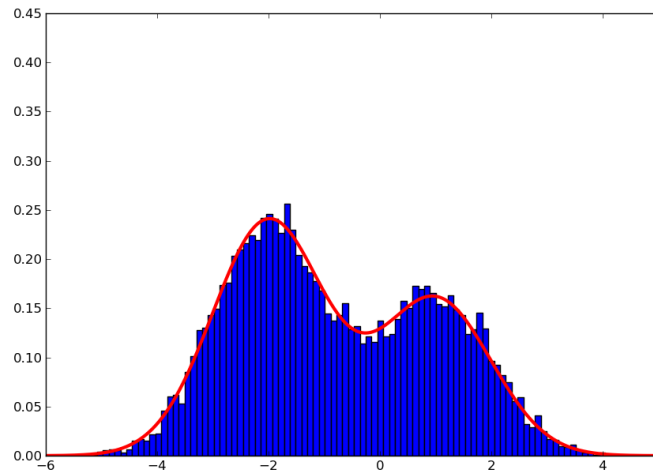
- 3ème possibilité : l'algorithme Langevin [Roberts et Rosenthal (1998)]

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

σ : facteur d'échelle

 Il existe plein d'autres possibilités. . .

Random walk avec une loi normale centrée réduite

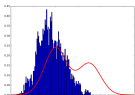
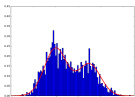
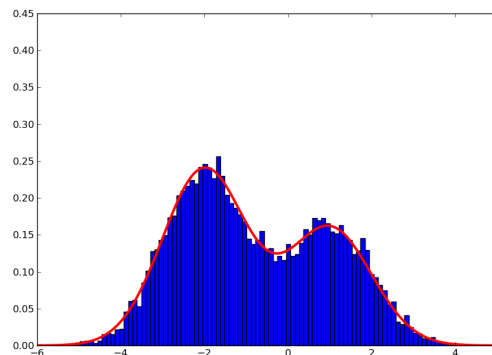


⚠ important pour la vitesse de convergence

Influence de l'étalement

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

$\sigma = 1$



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 :

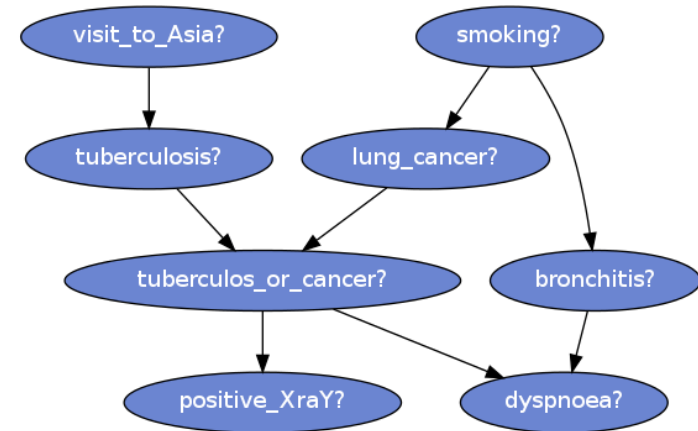
Partir de n'importe quelle valeur x_0

⚠ au départ l'échantillon ne suit pas $\tilde{\pi}(\cdot)$

⇒ 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



⚠ $x_t =$ vecteur à 8 valeurs !

Metropolis-Hastings par bloc

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

Précédemment :

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

Maintenant :

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

• stationnarité :

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

• 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

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

Rappel : stationnarité pour 1 variable

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

Stationnarité par bloc

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

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

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

Metropolis-Hastings par bloc : stationnarité

conséquences de la stationnarité par bloc :

$$\begin{aligned}
 & \int_{x_t^1} \int_{x_t^2} P(x_{t+1}^2 | x_{t+1}^1, x_t^2) \times P(x_{t+1}^1 | x_t^1, x_t^2) \tilde{\pi}(x_t^1, x_t^2) dx_t^1 dx_t^2 \\
 &= \int_{x_t^1} \int_{x_t^2} P(x_{t+1}^2 | x_{t+1}^1, x_t^2) \times P(x_{t+1}^1 | x_t^1, x_t^2) \tilde{\pi}(x_t^1 | x_t^2) \tilde{\pi}(x_t^2) dx_t^1 dx_t^2 \\
 &= \int_{x_t^2} \int_{x_t^1} P(x_{t+1}^2 | x_{t+1}^1, x_t^2) \times P(x_{t+1}^1 | x_t^1, x_t^2) \tilde{\pi}(x_t^1 | x_t^2) \tilde{\pi}(x_t^2) dx_t^1 dx_t^2 \\
 &= \int_{x_t^2} P(x_{t+1}^2 | x_{t+1}^1, x_t^2) \left[\int_{x_t^1} P(x_{t+1}^1 | x_t^1, x_t^2) \tilde{\pi}(x_t^1 | x_t^2) dx_t^1 \right] \tilde{\pi}(x_t^2) dx_t^2 \\
 &= \int_{x_t^2} P(x_{t+1}^2 | x_{t+1}^1, x_t^2) \tilde{\pi}(x_{t+1}^1 | x_t^2) \tilde{\pi}(x_t^2) dx_t^2 \quad (\text{stationnarité par bloc}) \\
 &= \int_{x_t^2} P(x_{t+1}^2 | x_{t+1}^1, x_t^2) \tilde{\pi}(x_t^2 | x_{t+1}^1) \tilde{\pi}(x_{t+1}^1) dx_t^2 \quad (\text{formule de Bayes}) \\
 &= \tilde{\pi}(x_{t+1}^2 | x_{t+1}^1) \tilde{\pi}(x_{t+1}^1) = \tilde{\pi}(x_{t+1}^1, x_{t+1}^2) \quad (\text{stationnarité par bloc})
 \end{aligned}$$

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

Échantillonneur de Gibbs

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

Conséquence :

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

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

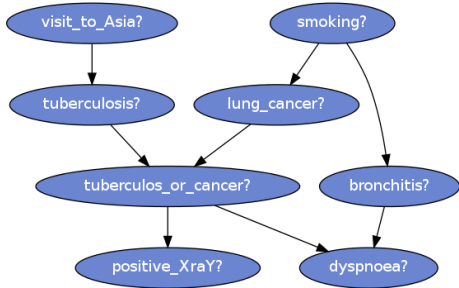
Échantillonneur de Gibbs

Algorithme

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

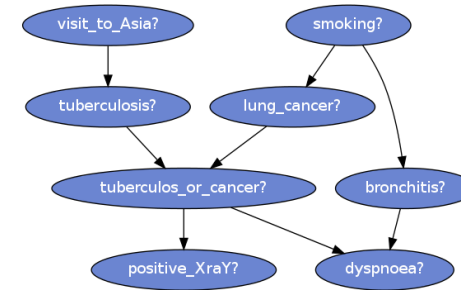
- 1 choisir une permutation $\sigma : \{1, \dots, n\} \mapsto \{1, \dots, n\}$
- 2 pour tout $i \in \{1, \dots, n\}$ faire :
 - a posons $y = (x_{t+1}^{\sigma(1)}, \dots, x_{t+1}^{\sigma(i-1)}, x_t^{\sigma(i+1)}, \dots, x_t^{\sigma(n)})$
 - b tirer $x_{t+1}^{\sigma(i)}$ selon la distribution $\tilde{\pi}(\cdot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

Asia et l'échantillonneur de Gibbs



- $\pi(\cdot)$ connu : c'est la distribution jointe du réseau bayésien
- nouvelle valeur de $B = \text{"bronchitis"}$?
 \Rightarrow échantillonner selon $P(B|y)$, avec y les valeurs de toutes les autres variables
- Si on note par les initiales minuscules les valeurs observées :
 $P(B|y) = P(B|vta, s, t, lc, toc, px, d)$

Asia et l'échantillonneur de Gibbs



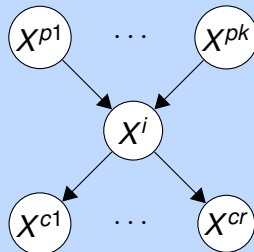
$$P(B|y) = \frac{P(B, y)}{P(y)} \propto P(B, y) = \text{vecteur de taille } |\text{bronchitis}|$$

$$P(B, y) = P(B, vta, s, t, lc, toc, px, d)$$

$$= P(vta)P(s)P(t|s)P(lc|s)P(toc|t, lc)P(B|s)P(px|toc)P(d|toc, B) \\ \propto P(B|s)P(d|toc, B)$$

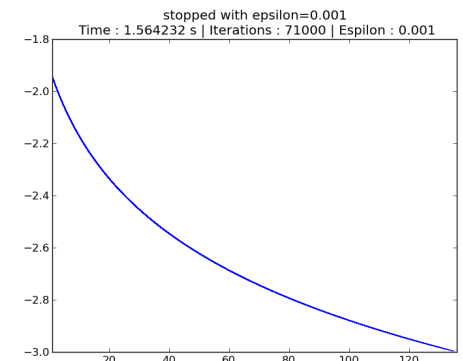
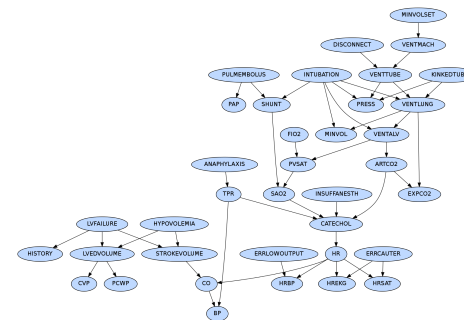
Échantillonneur de Gibbs et les réseaux bayésiens

Règle pour déterminer la valeur de x_{t+1}^i



- 1 extraire le vecteur $V = P(X^i|x^{p1}, \dots, x^{pk})$
- 2 extraire les vecteurs $V_j = P(x^{ci}|pa(X^{ci}), X^i)$
 où $pa(X^{ci}) = \text{valeur des parents de } X^{ci} \text{ sauf } X^i$
- 3 calculer la distribution $\Pi = V \times \prod_{i=1}^r V_j$ (produits tensoriels : terme à terme) et la normaliser
- 3 tirer x_{t+1}^i selon Π (distribution discrète)

Convergence de Gibbs sur le réseau "alarm"



- Ordonnées : $\log_{10}(\sum_i \text{distance Kullback Leibler}(P(X_t^i), P(X_{t+1}^i)))$
 $-3 \Rightarrow \sum_i \text{dist KL} = 10^{-3}$