**Introduction aux modèles graphiques** **2007/2008**

Cours 8 — 12 decembre 2007

*Enseignant: Francis Bach* *Scribe: Anne-Marie Tousch, Otakar Franc*

## 8.1 Rappel dernier cours

Nous avions vu comment faire de l'inférence exacte, avec une complexité $O(e^{\text{largeur arborescente}})$. Comment faire lorsque la largeur arborescente devient grande ? On fait de l'inférence approchée, et il existe deux types de méthodes :

1. par échantillonnage, que nous allons voir aujourd'hui,
2. variationnelles, au programme de la semaine prochaine.

## 8.2 Méthodes d'échantillonnage : introduction

Soit $X$ une v.a. de densité $p(x)$, et $f$ une fonction.

**But :** estimer $\mu = \mathbf{E}\, f(X) = \int p(x) f(x) dx$.

**Exemples :**

1. $f(x) = x$, $\mu$ est la moyenne
   $f(x) = x^2$, $\mu$ permet de calculer la variance
   Idée : calculer la marginale revient à calculer une espérance.
2. soit $X = (X_1, \ldots, X_n)$ multivariée discrète. Soit $f(x) = \delta(X_i = j)$. $\mathbf{E}\, f(x) = P(X_i = j)$.

**Méthode générale** Soit un échantillon $X_1, \ldots, X_n$ iid de loi $p(x)$. On considère l'estimateur suivant :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

Loi forte des grands nombres :

$$\mathbf{E}(f(x)) < \infty \Rightarrow \hat{\mu} \to \mu \text{ p.s.}$$

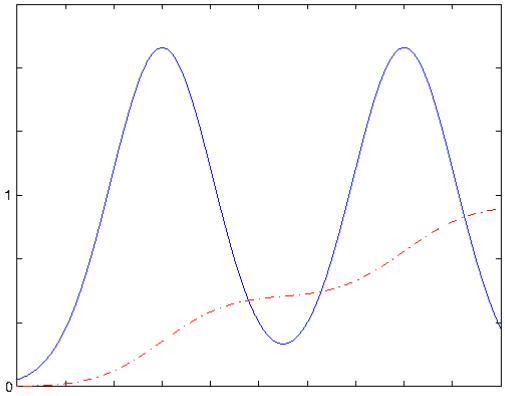



**Fig. 8.1 :**

Propriétés asymptotiques :

$$\begin{aligned} \mathbf{E}\,\hat{\mu} &= \frac{1}{n}\sum \mathbf{E}\, f(X_i) = \mathbf{E}\, f(X) = \mu \\ Var(\hat{\mu}) &= \frac{1}{n^2}\sum Var(f(X_i)) = \frac{1}{n} Var(f(X)). \end{aligned}$$

– Avantage : Indépendance par rapport à $n$.
– Inconvénient : $Var(f(X))$ est inconnu et potentiellement grand.
Dans les parties suivantes, nous allons voir les différentes méthodes d'échantillonnage.

## 8.3 Loi uniforme

$U(0,1)$ ou `rand` : nous n'allons pas le présenter !

## 8.4 Technique de la fonction de répartition

Soit $p(x)$ densité sur $\mathbb{R}$.

**Définition 8.1 (Fonction de répartition)**

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} p(t)dt$$

**Exemple :** voir figure 8.1.
Algorithme :

1. échantillonner sur $u \sim U[0,1]$,
2. $y = F^{-1}(u)$.

**Proposition 1** *$y$ a pour loi $p(x)$.*

**Démonstration**

$$\begin{aligned} P(y \leq x) &= P(F(y) \leq F(x)) \\ &= P(u \leq F(x)) \\ &= F(x) \end{aligned}$$

■

- Avantage : facile
- Inconvénient : il faut calculer et inverser $F(x)$.

## 8.5 Échantillonnage par rejet

### 8.5.1 Cadre

$p(x) = \frac{\tilde{p}(x)}{Z_p}$, $Z_p$ est "inconnu" ($Z_p = \sum \tilde{p}$, difficile à calculer).

### 8.5.2 Rejet

On suppose connue $q(x)$ densité tq. :

1. on sait échantillonner $q(x)$,
2. $\exists k > 0,\ \forall x,\ k \cdot q(x) \geq \tilde{p}(x)$.

### 8.5.3 Algorithme

1. Echantillonner $x \sim q(x)$
2. Accepter $x$ avec probabilité $\frac{\tilde{p}(x)}{k \cdot q(x)}$.
3. Si rejet, recommencer en 1.

**Exemple 1D :** cf. figure 8.2
En pratique : on échantillonne $u \sim U[0,1]$. Si $u \leq \frac{\tilde{p}(x)}{k \cdot q(x)}$ accepte, sinon rejette.

**Proposition 2** *La loi $p(y)$ recherchée est la loi $p(x|accepté)$.*

**Démonstration**

$$\begin{aligned} p(x = y|\text{accepté}) &\propto p(x = y\&\text{accepté}) \\ &\propto q(y)\frac{\tilde{p}(y)}{k \cdot q(y)} = \frac{\tilde{p}(y)}{k} \\ \Rightarrow p(x = y|\text{accepté}) &= p(y). \end{aligned}$$

■

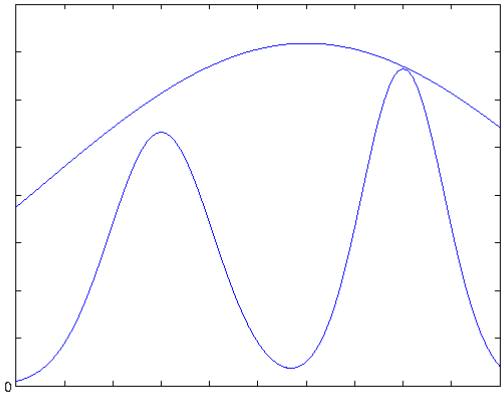



**Fig. 8.2 :**

**Proposition 3** *La probabilité d'acceptation vaut :*

$$\begin{aligned} p(accept) &= \mathop{\mathbf{E}}_{q(x)} \frac{\tilde{p}(x)}{k \cdot q(x)} = \int q(x) \frac{\tilde{p}(x)}{k \cdot q(x)} dx \\ &= \frac{1}{k} \int \tilde{p}(x) = \frac{Z_p}{k} \end{aligned}$$

Inconvénients de la méthode :
- Calcul de $k$ : $q(x)$ doit avoir une "queue lourde" (ex. Cauchy, exponentielle).
- $k$ grand, problématique si $X$ a une grande dimension. Un cas simple en une dimension, pour les densités log-concace est la méthode *adaptive rejection sampling* (voir poly).

Pour les grandes dimensions, on utilisera la méthode correspondant au paragraphe suivant...

## 8.6 Echantillonnage d'importance

Cette méthode ne donne pas un échantillonnage de $p$, mais elle permet néanmoins de calculer l'espérance.

l'idée génréral est la suivante :

$$\begin{aligned} \int f(x)p(x)dx &= \int \frac{f(x)p(x)}{q(x)} q(x)dx \\ p(x) &= \frac{\tilde{p}(x)}{Z_p} \end{aligned}$$

On calcule l'estimateur sur un échantillon $x_1, \ldots, x_n$ iid suivant $q(x)$ (qui est supposée facile à échantilloner) :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} f(x_i) \frac{p(x_i)}{q(x_i)}.$$

On a donc :

$$\begin{aligned}\mathbf{E}(\hat{\mu}) &= \frac{1}{n}\sum \int f(x)\frac{p(x)}{q(x)}q(x)dx = \int f(x)p(x)dx \\ Var(\hat{\mu}) &= \frac{1}{n}Var_{q(x)}\left(\frac{f(x)p(x)}{q(x)}\right)\end{aligned}$$

**Lemme 8.2** *Si* $\forall x,\ |f(x)| \leq M$,

$$Var(\hat{\mu}) \leq \frac{M^2}{n}\int \frac{p(x)^2}{q(x)}dx.$$

**Démonstration**

$$\begin{aligned}Var(\hat{\mu}) = &\ \tfrac{1}{n}Var_{q(x)}\left(\tfrac{f(x)p(x)}{q(x)}\right) \\ \leq &\ \tfrac{1}{n}\int \tfrac{f(x)^2p(x)^2}{q(x)^2}q(x)dx \\ \leq &\ \tfrac{M^2}{n}\int \tfrac{p(x)^2}{q(x)}dx.\end{aligned}$$

■

$$\begin{aligned}\int \frac{p(x)^2}{q(x)}dx &= \int \frac{p^2(x) - 2p(x)q(x) + q^2(x)}{q(x)}dx + \int \frac{2p(x)q(x) - q^2(x)}{q(x)}dx \\ &= \underbrace{\int \frac{(p(x)-q(x))^2}{q(x)}dx}_{\text{Divergence du } \chi^2 \text{ entre } p \text{ et } q.} + 1\end{aligned}$$

En conséquence, la méthode est correcte lorsque $p$ reste proche de $q$ (en particulier, $q$ a de la masse partout o'u $p$ en a, et les queues de $q$ sont plus lourdes que celles de $p$).

Cadre général (où on ne connait que $q$ à une constante près) : Soit $p(x) = \frac{\tilde{p}(x)}{Z_p}$, $q(x) = \frac{\tilde{q}(x)}{Z_q}$. On prend un échantillon iid $X_1, \ldots, X_n \sim q$.

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i)\frac{\tilde{p}(x_i)}{\tilde{q}(x_i)} = \frac{Z_p}{Z_q}\hat{\mu} \xrightarrow{esp.} \frac{Z_p}{Z_q}\int f(x)p(x)dx.$$

Dans le cas où $f = 1$, on a

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\tilde{p}(x_i)}{\tilde{q}(x_i)} \xrightarrow{esp.} \frac{Z_p}{Z_q}$$

Dans le cas général, on estime l'espérance :

$$\hat{\mu}_n = \frac{\frac{1}{n}\sum_{i=1}^{n} f(x_i)\frac{\tilde{p}(x_i)}{\tilde{q}(x_i)}}{\frac{1}{n}\sum_{i=1}^{n} \frac{\tilde{p}(x_i)}{\tilde{q}(x_i)}}$$

**Proposition 4**

$$\mathbf{E}\,\|\hat{\mu}_n - \mu\|^2 = O\left(\frac{M^2}{n}\int\frac{p^2}{q}\right).$$

## 8.7 Markov Chain Monte Carlo (MCMC)

**Cadre** Soit $x \in X$, $X$ fini (mais très grande dimension).
On construit une chaîne de Markov : $X_0, X_1, \ldots$ tq. sa densité $q_t(x) = p(X_t = x)$ converge vers $p(x)$.

### 8.7.1 Rappel sur chaîne de Markov

**Définition 8.3 (MC homogène en temps)**

$$\begin{aligned}\forall t \geq 0\ \forall (x,y) \in X \qquad & p(X_{t+1} = y \mid X_t = x, X_{t-1}, \ldots, X_0)\\ & = p(X_{t+1} = y \mid X_t = x)\\ & = p(X_1 = y \mid X_0 = x)\\ & = S(x,y)\end{aligned}$$

**Définition 8.4 (matrice de transition)** *$S(x,y)$ est une matrice de transition.*

*Si on note $k = Card(X)$, alors :*
- $S \in \mathbb{R}^{k\times k}$
- $S \geq 0$
- $S1 = 1$. *(la somme d'une colonne vaut 1)*

Considérons les densités suivantes :
$q_0 \in \mathbb{R}^k$, $q_0(x) = P(X_0 = x), q_t(x) = P(X_t = x)$
En utilisant la notion de la matrice de transition :

$$\begin{aligned}q_{t+1}(x) &= P(X_{t+1} = x)\\ &= \sum_y P(X_{t+1} = x \mid X_t = y)P(X_t = y)\\ &= \sum_y S(y,x)q_t(y)\\ &= (S^T q_t)(x)\end{aligned}$$

On en déroule :

**Proposition 5**

$$\boxed{q_t = (S^T)^t q_0}$$

**Définition 8.5 (Loi stationnaire)** *Loi $\Pi(x)$ est stationnaire si :*
- $\Pi \geq 0$

– $\Pi 1 = 1$.

$$\begin{aligned} &\Leftrightarrow && \Pi = S^T \Pi \\ &\Leftrightarrow && \Pi = \textstyle\sum_x \Pi(x) S(x,y) \end{aligned}$$

**Théorème 8.6 (Perron-Frobenius)** *Si* $S \geq 0, S1 = 1$,
*alors* 1 *est la plus grande valeur propre associée à un vecteur* $v \geq 0$,
*c'est-à-dire,* $\exists v \geq 0, S^T v = v$

**Corollaire 8.7** $\forall S, \exists\ \Pi stationnaire$

On ne sait rien sur la unicité (chaine irréductible) ou si ca va converger vers $\Pi$ (chaine apériodique). Voir P. Bremaud, Markov Chains Texts In Applied Mathematics, Springer, 2001.

**Théorème 8.8** *Si* $\forall (x,y)\ ,\ S(x,y) > 0$, *alors :*
– $\exists!$ *loi stationnaire* $\Pi$
– $\forall\ q_0,\ q_t\ \rightarrow \Pi$ *pour* $t \rightarrow \infty$.

**But des méthodes MCMC** Etant donné $p(x)$, il faut définir/trouver $S(x,y)$ tel que :

1. $\sum_x S(x,y)p(x) = p(y)$
2. $\forall\ (x,y),\ S(x,y) > 0$.

**Lemme 8.9** *Si* $\forall\ (x,y)\ p(x)S(x,y) = p(y)S(y,x)$, *alors* $p$ *est stationnaire pour* $S$.

**Démonstration** $\sum_x S(x,y)p(x) = \sum_x p(y)S(y,x) = p(y) \underbrace{\sum_x S(y,x)}_{1} = p(y)$ ■

**Remarque :** L'équation $p(x)S(x,y) = p(y)S(y,x)$ du lemme précédent s'appelle bilan détaillé (*detailed balance* en anglais). Si l'équation "detailed balance" est vérifiée, alors la chaîne de Markov est dite réversible.

### 8.7.2 Transition de Metropolis-Hastings (MH)

1. Echantillonnage $z \sim T(x,z)$
2. avec probabilité $A(x,z)$ :
   – Si l'on accepte, alors $y = z$,
   – si l'on rejete, alors $y = x$.

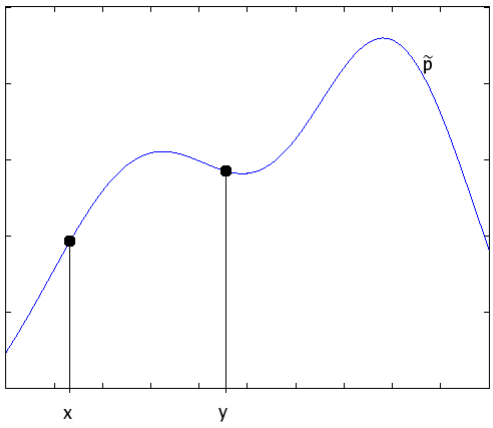



**FIG. 8.3 :** au niveau de $x$, $A(x_t, x_{t+1})$ vaudra 1, et au niveau de $y$ il vaudra $\frac{\tilde{p}(x_{t+1})}{\tilde{p}(x_t)}$.

**Proposition 6** *Le noyau associé est tel que :*

$$\forall y \neq x \; S(x,y) = T(x,y)A(x,y). \tag{8.1}$$

*Dans le cas de Metropolis-Hastings, on définit $A(x,y)$ de la manière suivante :*

$$A(x,y) = min\left(1, \frac{\tilde{p}(y)T(y,x)}{\tilde{p}(x)T(x,y)}\right) \tag{8.2}$$

*Si $T$ est symétrique, alors :*

$$A(x,y) = min\left(1, \frac{\tilde{p}(y)}{\tilde{p}(x)}\right) \tag{8.3}$$

**Vérification :** Pour la transition de Metropolis-Hastings :

**Proposition 7**

$$\begin{gathered}\forall (x,y), p(x)S(x,y) = p(y)S(x,y) \\ \Leftrightarrow \\ \forall x \neq y, p(x)S(x,y) = p(y)T(x,y)min\left(1, \frac{\tilde{p}(y)T(y,x)}{\tilde{p}(x)T(x,y)}\right)\end{gathered}$$

**Démonstration**

$$\begin{aligned} p(x)S(x,y) &= p(x)T(x,y)min\left(1, \frac{\tilde{p}(y)T(y,x)}{\tilde{p}(x)T(x,y)}\right) \\ &= min\left(p(y)T(x,y),\ p(x)T(y,x)\right) \\ &= p(y)S(x,y) \end{aligned}$$

■

### 8.7.3 Algorithme de MCMC/MH

1. Echantillonner $X_0$ arbitrairement
2. For $t = 1$ to ...
   (a) echantilloner y avec probabilité $T(X_t, y)$
   (b) – accepter avec probabilité $A(X_t, y) \Longrightarrow X_{t+1} = y$
   – sinon refuser $\Longrightarrow X_{t+1} = x$

**Problèmes associés :**

1. Vitesse de convergence dépend de la deuxième valeur propre de $S$
2. Quand faut-il arrêter ?
   – réponse théorique existantes mais difficiles à mettre en oeuvre
   – réponse pratique :
   (a) Burn-in période (par exemple 10000)
   (b) Garder un échantillon tous les $N$ donné (par exemple 100) échantillons.

## 8.8 Gibbs sampling

L'échantillonage de Gibbs est particuliìeremet adapté aux modèles graphiques. Soit
– $X = (X_1, \ldots, X_n)$
– $p(x) = \frac{\tilde{p}(x)}{Z}$

**Algorithme**

1. $X^0 = (X_1^0, \ldots, X_n^0)$ arbitraire
2. For $t = 1$ to ...
   – $X_1^{t+1} = p(X_1 \mid X_2^t, \ldots, X_n^t)$
   – $X_2^{t+1} = p(X_2 \mid X_1^{t+1}, X_3^t, \ldots, X_n^t)$
   – ...
   – $X_n^{t+1} = p(X_n \mid X_1^{t+1}, \ldots, X_{n-1}^{t+1})$

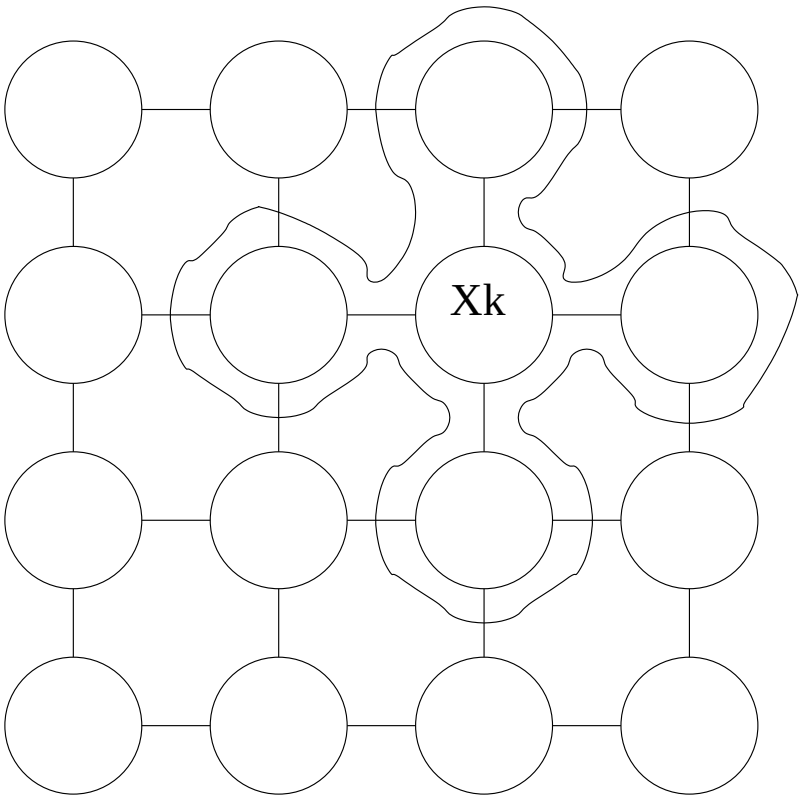
**Intérêt** par rapport aux modèles graphiques : $p(x_k \mid X_{reste})$ simple pour que ça marche

**Définition 8.10 (Couverture de Markov)** *La couverture de Markov est le plus petit ensemble $S$ tel que $p(x_k \mid X_{reste}) = p(x_k \mid x_S)$.*

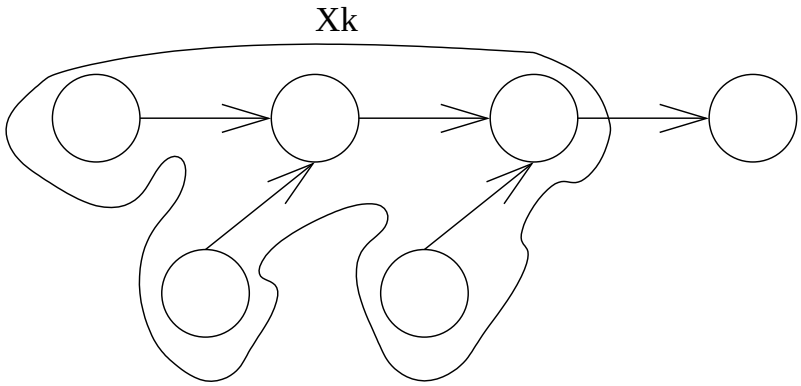
**Couverture pour graphes non orientés :** si $p(x)$ se factorise dans $G$, alors la couverture de Markov du sommet $k$ est composée des voisins de $k$.

**Couverture pour graphes orientés :** $p(x) \in \mathbb{L}(G)$, G orienté :
Couverture de $X_k$ est formée par :
– parents de $X_k$
– enfants de $X_k$
– parents des enfants de $X_k$

**Exemples de la couverture de Markov :** $p(x_k) \in \mathbb{L}(G)$, G non orienté




Couverture de Markov d'un graphe non-orienté




Couverture de Markov d'un graphe orienté

(C'est le même cas que si on se ramène au graphe moralisé)

### 8.8.1 Gibbs randomisé

1. (comme pour le Gibbs sampling)
2. For $t = 1$ to ...
   (a) Choisir $k$ aléatoirement

(b) On remplace $X_k^{t+1} \sim p(x_k \mid X_{reste}^t)$

**Proposition 8** *Gibbs randomisé est une instance de Metropolis-Hastings avec* $A(x, y) = 1$

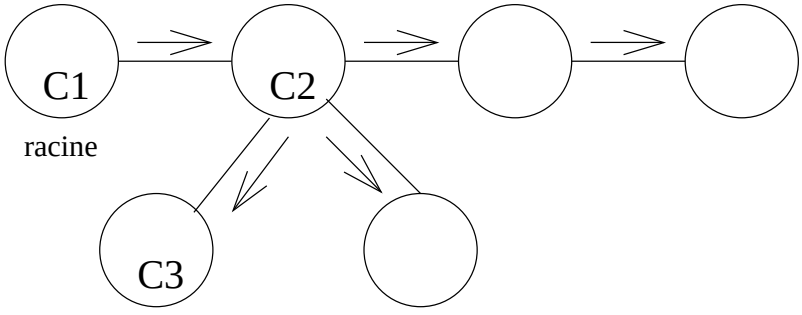
## 8.9 Modèles graphiques

Pourquoi échantillonner ?
– inférence
– exploration
Soit $p(x) \in \mathbb{L}(G)$, comment échantillonner x ?

1. cas général : Gibbs
2. DAG G - sans observation
   Il suit l'ordre topologique $1..n$

   **algorithm** For $i = 1$ to $n$ échantillonner $X_i \mid X_{\Pi_i}$
3. DAG avec observations

   **algorithm**
   (a) Echantillonner le DAG sans observations
   (b) On ne garde que les échantillonnes compatibles
4. Arbre de jonctions



Arbre de jonctions

   **algorithm**
   (a) Choisir une racine
   (b) Propager l'échantillon

**Conclusion :** Méthodes d'échantillonage convergent lentement, par rapport aux méthodes de variationnelles qui sont plus rapide mais ne convergent pas exactement vers le résultat exact.