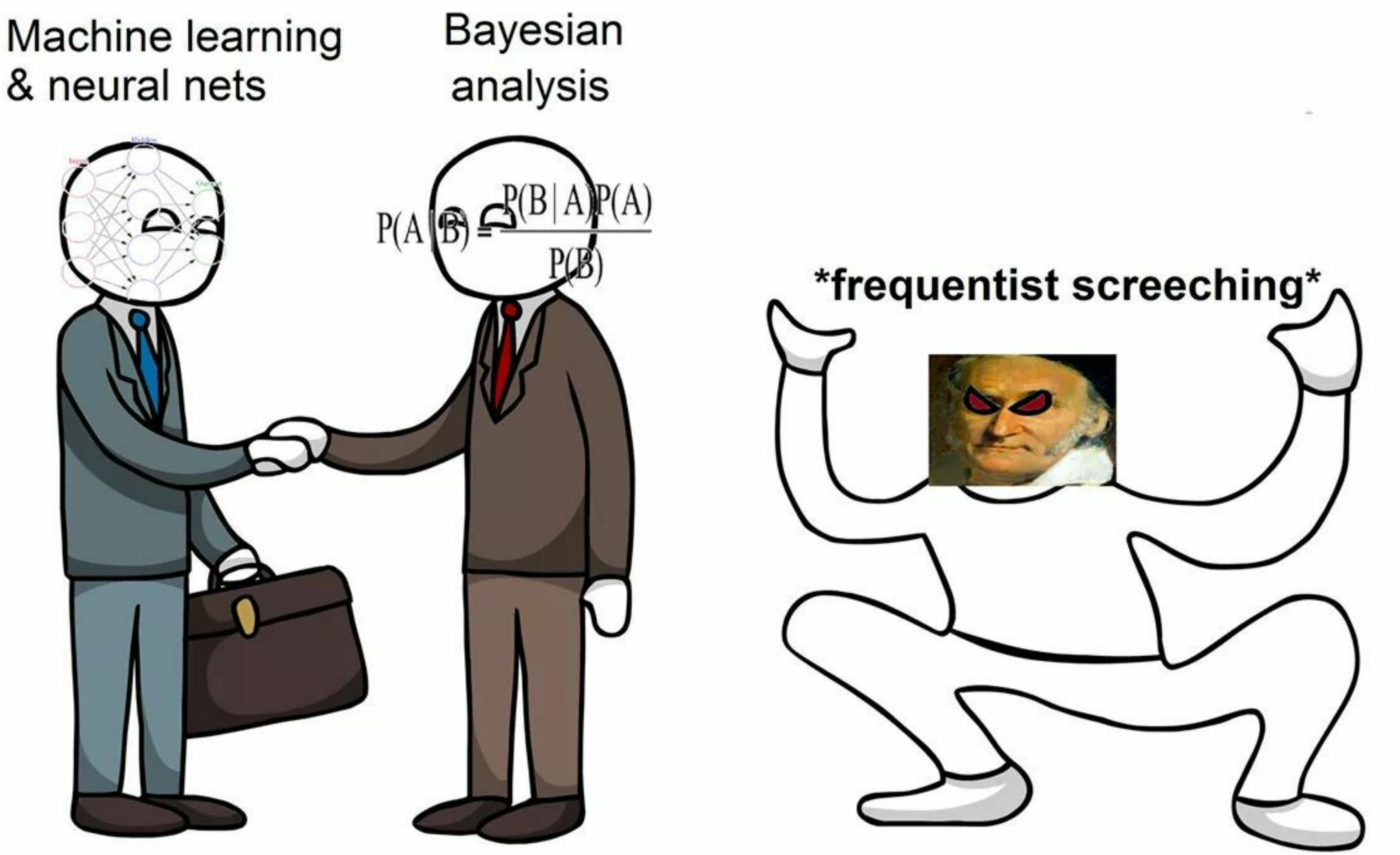
Apprentissage Bayésien de graphes dirigés acycliques appliqué à l'identification du modèle en apprentissage machine

Patrick Dallaire

Centre de Recherche en Données Massive
Université Laval

2 février 2018


Machine learning
& neural nets
Bayesian
analysis
$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
frequentist screeching

Plan de présentation

- Introduction à l'apprentissage Bayesien
- Présentation du processus des chefs Indiens
- Application à l'apprentissage de réseaux Bayésiens
- Potentiel d'application pour apprendre la structure d'un réseau de neurones?

Plan de présentation

- **Introduction à l'apprentissage Bayesien**
- Présentation du processus des chefs Indiens
- Application à l'apprentissage de réseaux Bayésiens
- Potentiel d'application pour apprendre la structure d'un réseau de neurones?

L'approche Bayésienne

L'approche Bayésienne standard:

1. Définir un **modèle** statistique

$$p(y_i \mid x_i, \boldsymbol{\theta})$$

2. Formuler nos **a priori** sous forme de probabilité

$$p(\theta_j)$$

3. **Mise à jour** avec la règle de Bayes

$$p(\boldsymbol{\theta} \mid X, Y) = \frac{p(Y \mid \boldsymbol{\theta}, X)p(\boldsymbol{\theta})}{p(Y)} \propto p(Y \mid \boldsymbol{\theta}, X)p(\boldsymbol{\theta})$$

Exemple - Régression Linéaire

L'approche Bayésienne standard:

1. Définir un **modèle** statistique

$$p(y_i \mid x_i, \boldsymbol{\theta}) = \mathcal{N}(\theta_1 + \theta_2 x_i, 1)$$

2. Formuler nos **a priori** sous forme de probabilité

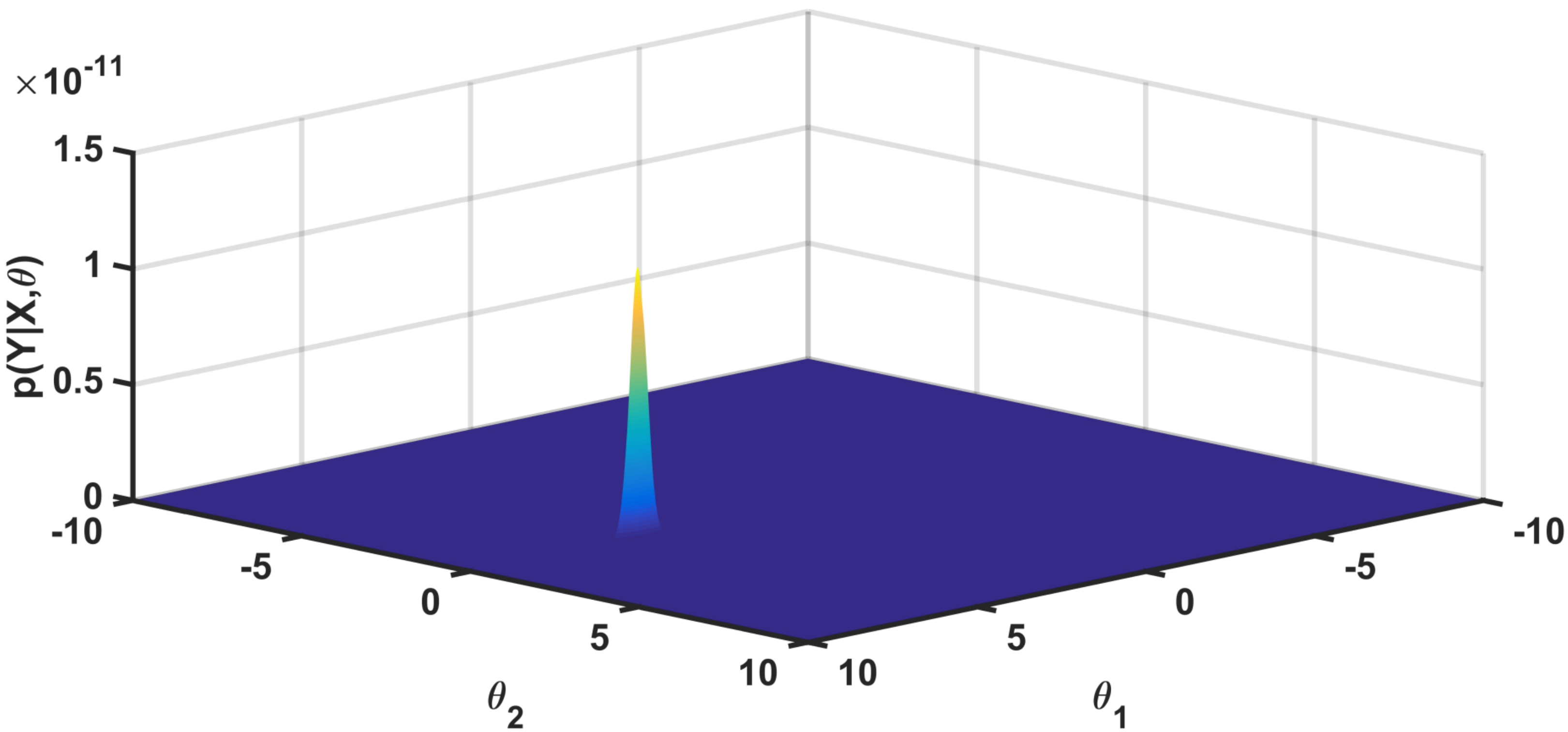
$$p(\theta_j) = \text{Laplace}(0, 1)$$

3. **Mise à jour** avec la règle de Bayes

$$p(\boldsymbol{\theta} \mid X, Y) = \frac{p(Y \mid \boldsymbol{\theta}, X)p(\boldsymbol{\theta})}{p(Y)} \propto p(Y \mid \boldsymbol{\theta}, X)p(\boldsymbol{\theta})$$

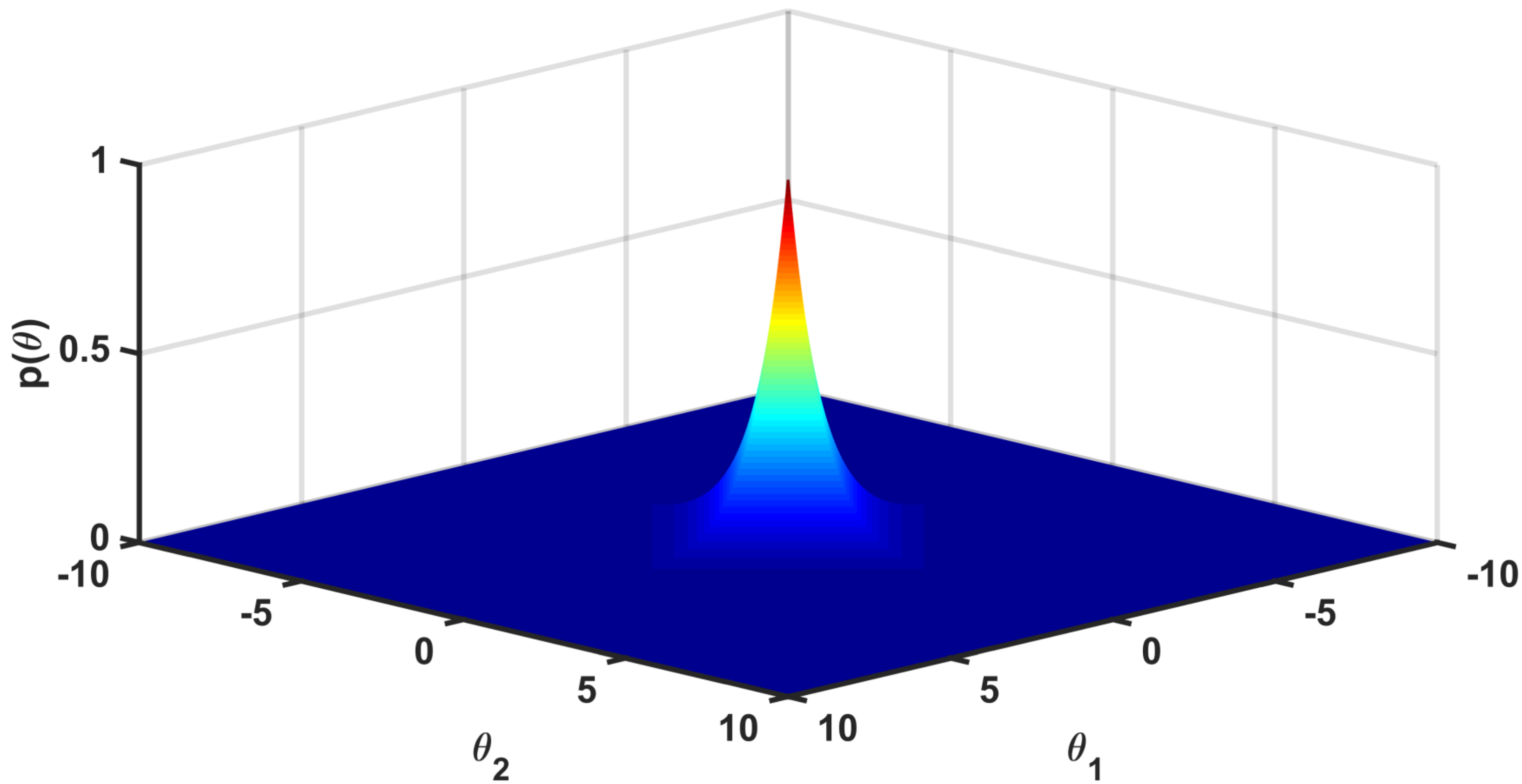
Fonction de vraisemblance

$$p(Y \mid X, \boldsymbol{\theta}) = \prod_{i=1}^{N} \mathcal{N}(y_i \mid f_{\boldsymbol{\theta}(x_i)}, 1)$$

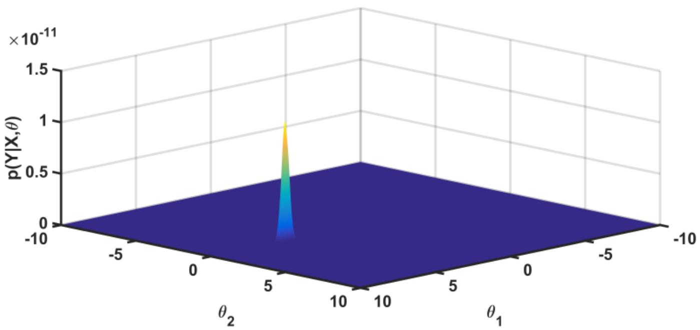


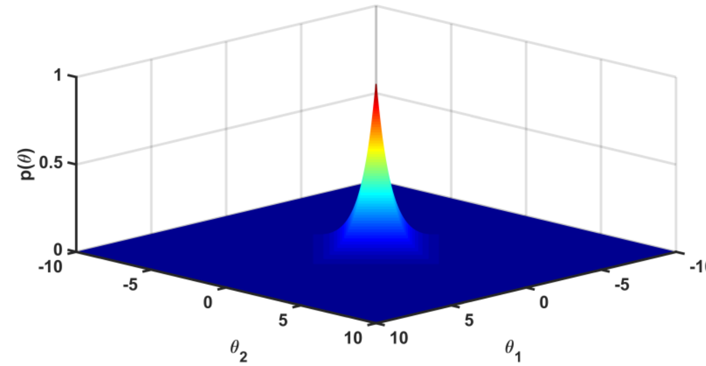
La probabilité *a priori*

$$p(\boldsymbol{\theta}) = \prod_j \mathsf{Laplace}(\theta_j \mid 0, 1) = \prod_j \frac{1}{2} \exp(-|\theta_j|)$$

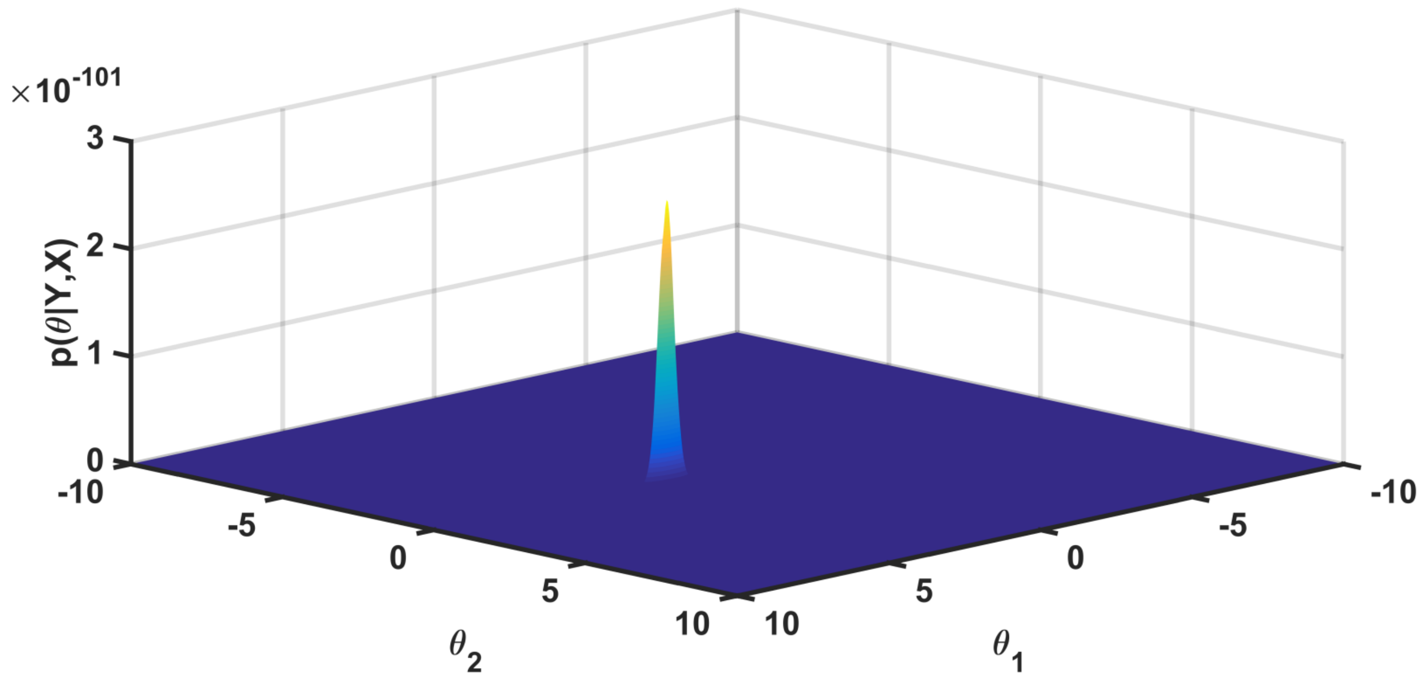


La probabilité *a posteriori*



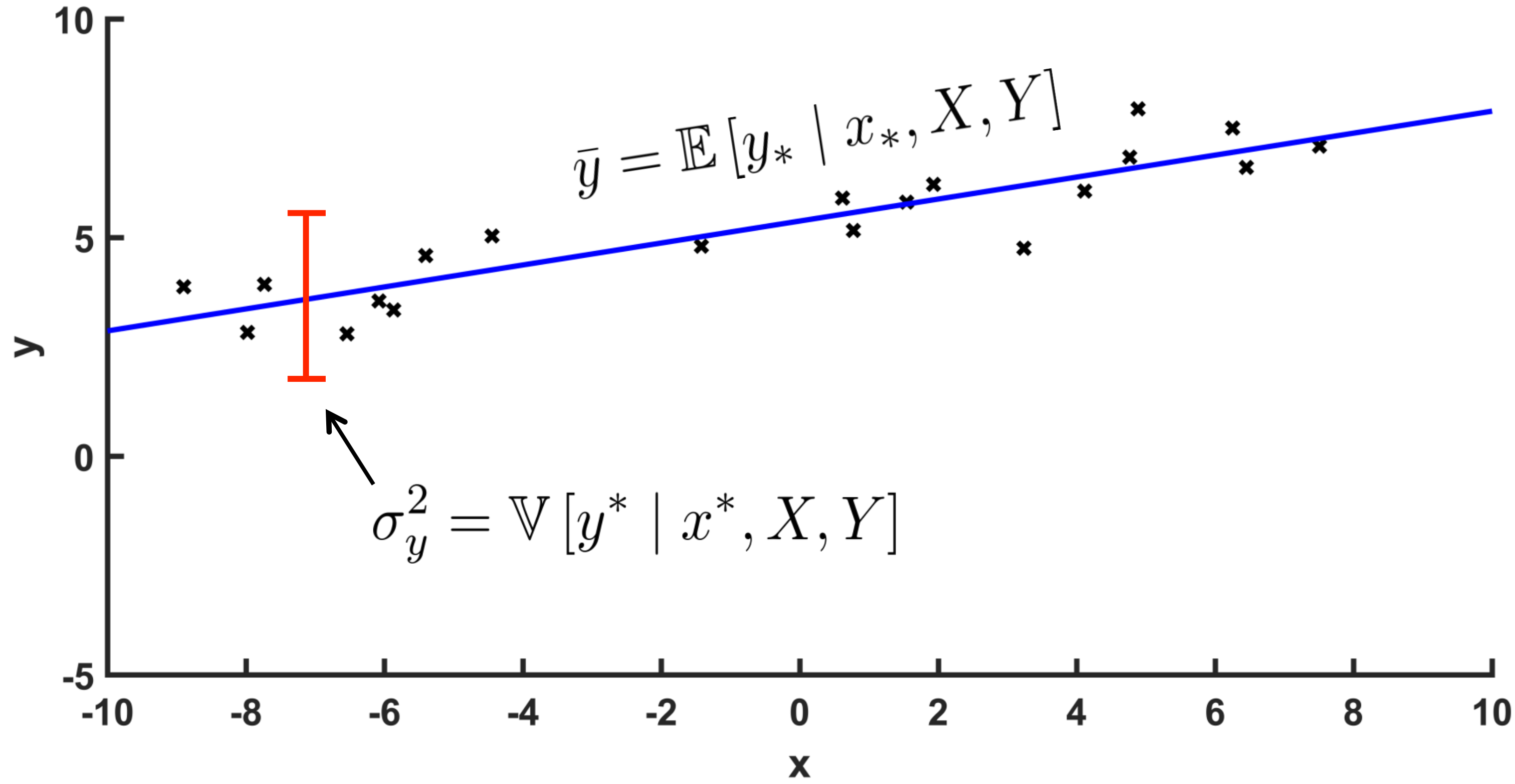


$$p(\boldsymbol{\theta}|X,Y) \propto p(Y|\boldsymbol{\theta},X)p(\boldsymbol{\theta})$$



Distribution prédictive a posteriori

$$p(y_* \mid x_*, X, Y) = \int p(y_* \mid x_*, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid X, Y) \mathrm{d}\boldsymbol{\theta}$$



Intégration Monte Carlo

$$p(y_* \mid x_*, X, Y) = \int p(y_* \mid x_*, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid X, Y) \mathrm{d}\boldsymbol{\theta}$$

L'intégral devient une sommation

Échantillonner le posterior

Comment échantillonner le posterior?

Markov Chain Monte Carlo (MCMC)

Metropolis-Hastings MCMC

Objectif : échantillonner une distribution $p(\boldsymbol{\theta})$

1) Initialiser aléatoirement $\widetilde{\boldsymbol{\theta}}_0$ et fixer à $t = 0$

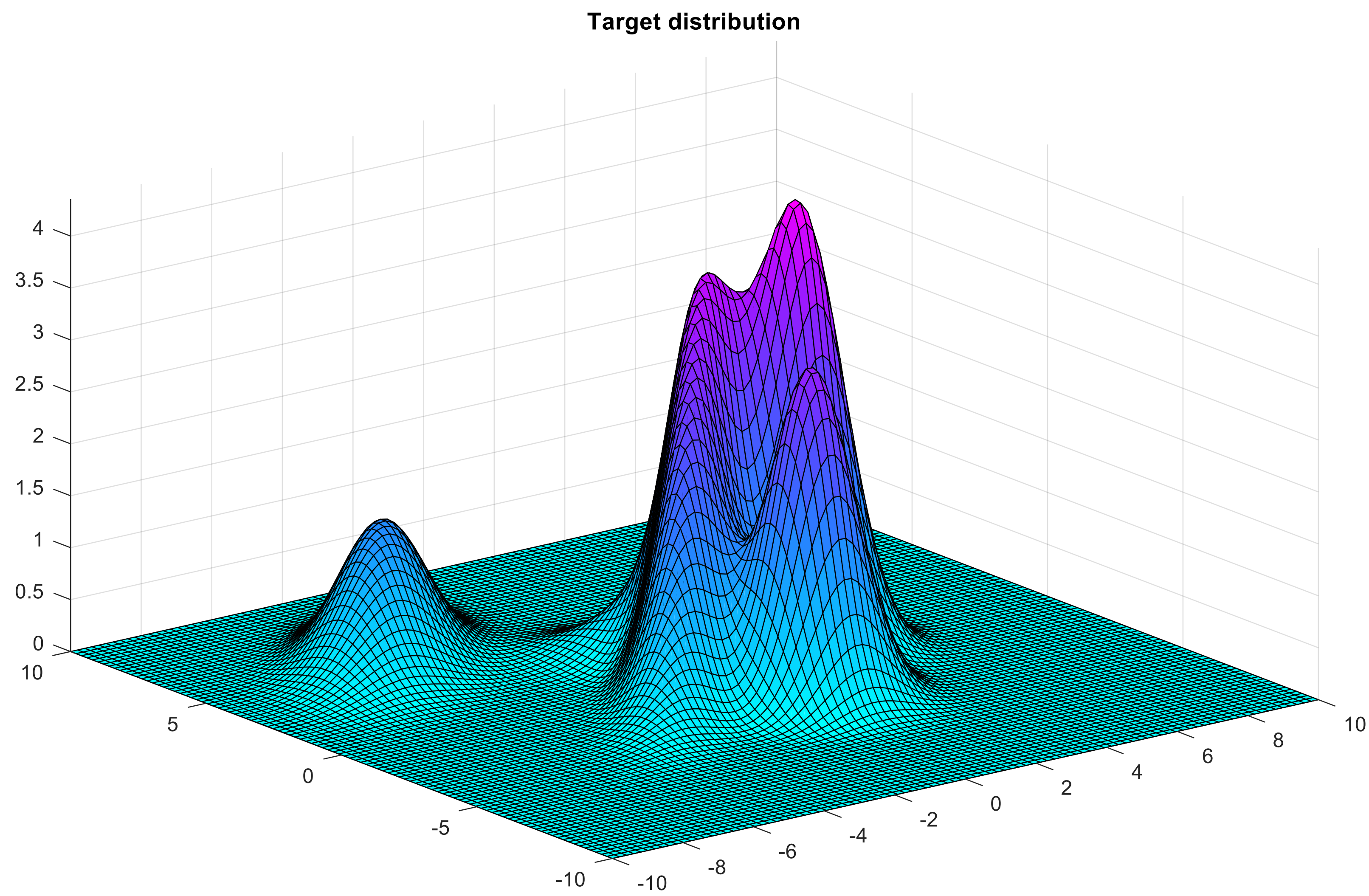
2) Générer un candidat $\boldsymbol{\theta}'$ à partir du *proposal* $g(\boldsymbol{\theta}' \mid \widetilde{\boldsymbol{\theta}}_t)$

3) Calculer la probabilité d'acceptation :

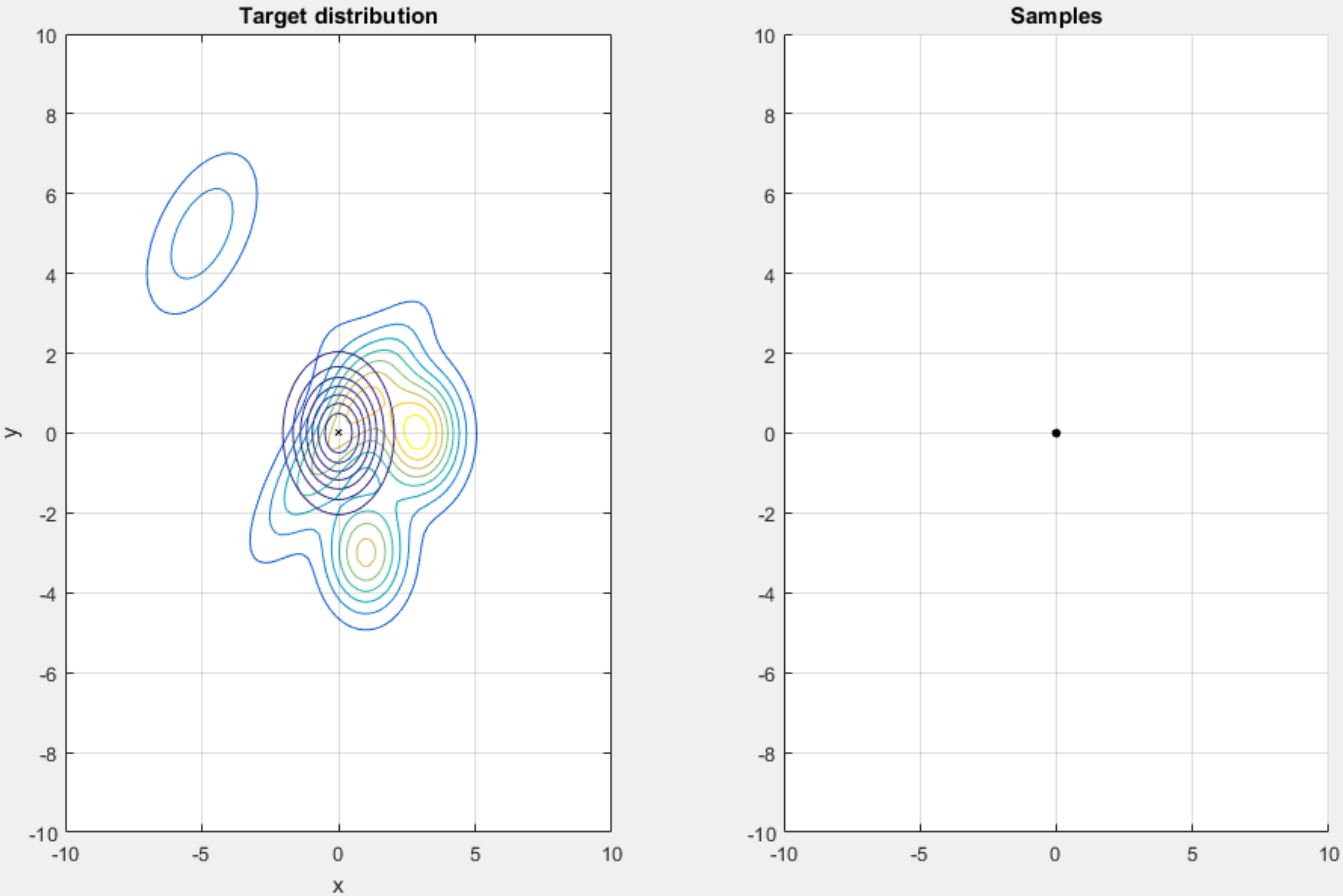
$$A(\boldsymbol{\theta}' \mid \widetilde{\boldsymbol{\theta}}_t) = \min\left(1, \frac{p(\boldsymbol{\theta}')g(\widetilde{\boldsymbol{\theta}}_t \mid \boldsymbol{\theta}')}{p(\widetilde{\boldsymbol{\theta}}_t)\, g(\boldsymbol{\theta}' \mid \widetilde{\boldsymbol{\theta}}_t)}\right)$$

4) Avec probabilité $A(\boldsymbol{\theta}' \mid \widetilde{\boldsymbol{\theta}}_t)$, accepter et $\widetilde{\boldsymbol{\theta}}_{t+1} = \boldsymbol{\theta}'$,
sinon rejeter et $\widetilde{\boldsymbol{\theta}}_{t+1} = \widetilde{\boldsymbol{\theta}}_t$

Exemple - Metropolis-Hastings



Exemple - Metropolis-Hastings



Plan de présentation

- Introduction à l'apprentissage Bayesien
- **Présentation du processus des chefs Indiens**
- Application à l'apprentissage de réseaux Bayésiens
- Potentiel d'application pour apprendre la structure d'un réseau de neurones?

Apprentissage Bayesien du modèle

Définir un modèle probabiliste pour les observations :

$$p(y_i \mid x_i, \boldsymbol{\theta}, M)$$

Définir le prior sur les paramètres et les modèles :

$$p(\boldsymbol{\theta}, M) = p(\boldsymbol{\theta} \mid M)p(M)$$

Procéder à l'inférence :

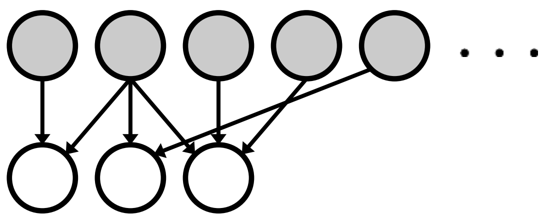
$$p(\boldsymbol{\theta}, M \mid X, Y) = \frac{p(Y \mid \boldsymbol{\theta}, X)p(\boldsymbol{\theta} \mid M)p(M)}{\iint p(Y \mid \boldsymbol{\theta}, X)p(\boldsymbol{\theta} \mid M)p(M)\mathrm{d}\boldsymbol{\theta}\mathrm{d}M}$$

Faire une prédiction à partir de l'*a posteriori* :

$$p(y_* \mid x_*, X, Y) = \iint p(y_* \mid x_*, \boldsymbol{\theta}, M)p(\boldsymbol{\theta}, M \mid X, Y)\mathrm{d}\boldsymbol{\theta}$$

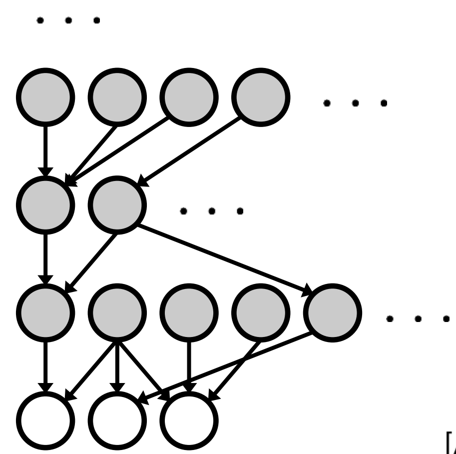
Distribution sur les DAG

Processus du buffet Indien (IBP)



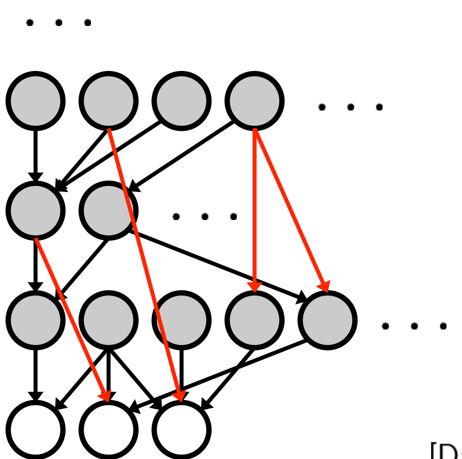
[Woods et al., 2006]

IBP en cascade (CIBP)



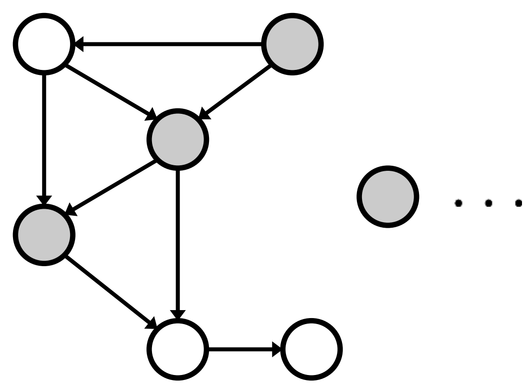
[Adams et al., 2010]

CIBP étendu (ECIBP)



[Dallaire et al., 2014]

Processus des chefs Indiens (ICP)



[*submitted*]

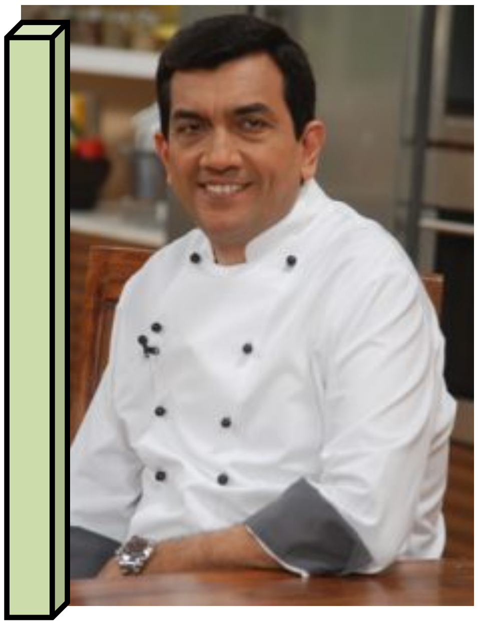
Top 3 des chefs Indiens

Sanjeev Kapoor

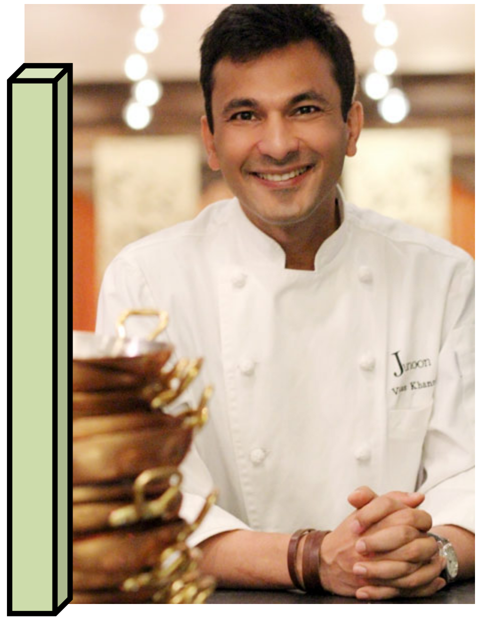
Vikas Khanna

Ranveer Brar

Processus des chefs Indiens

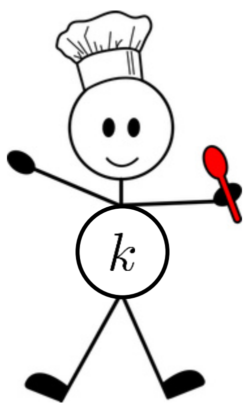


Sanjeev Kapoor



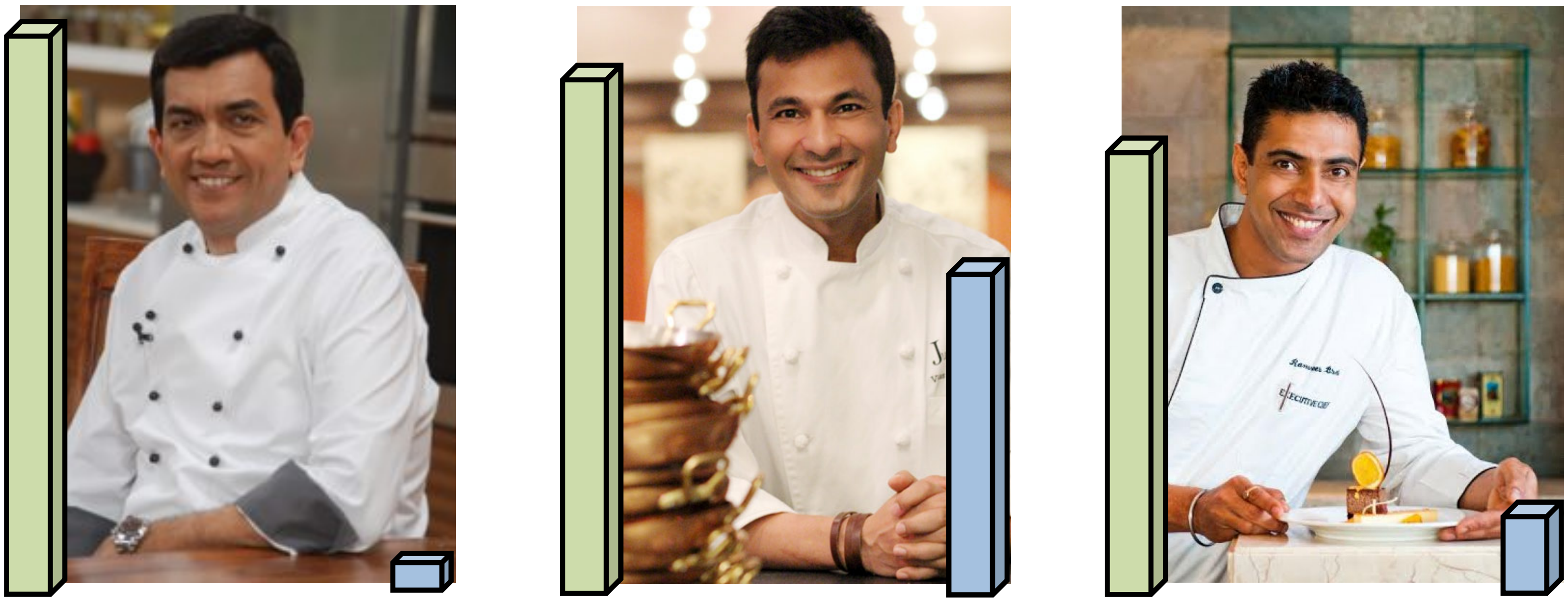
Vikas Khanna

Ranveer Brar

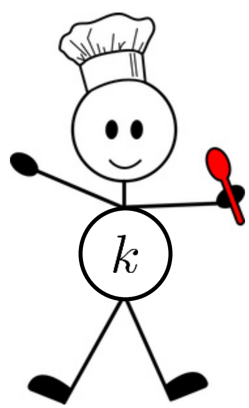


Réputation θ_k

Processus des chefs Indiens



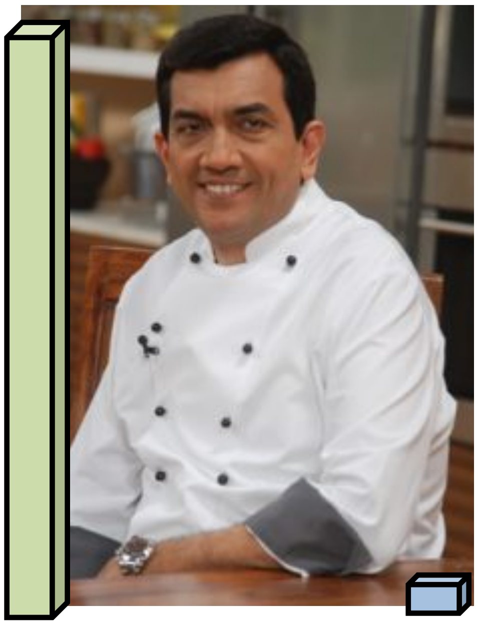
Sanjeev Kapoor **Vikas Khanna** **Ranveer Brar**



Réputation θ_k

Popularité π_k

Processus des chefs Indiens

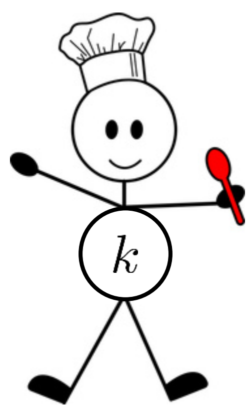


Sanjeev Kapoor

Vikas Khanna

Ranveer Brar

$K *$



Réputation θ_k

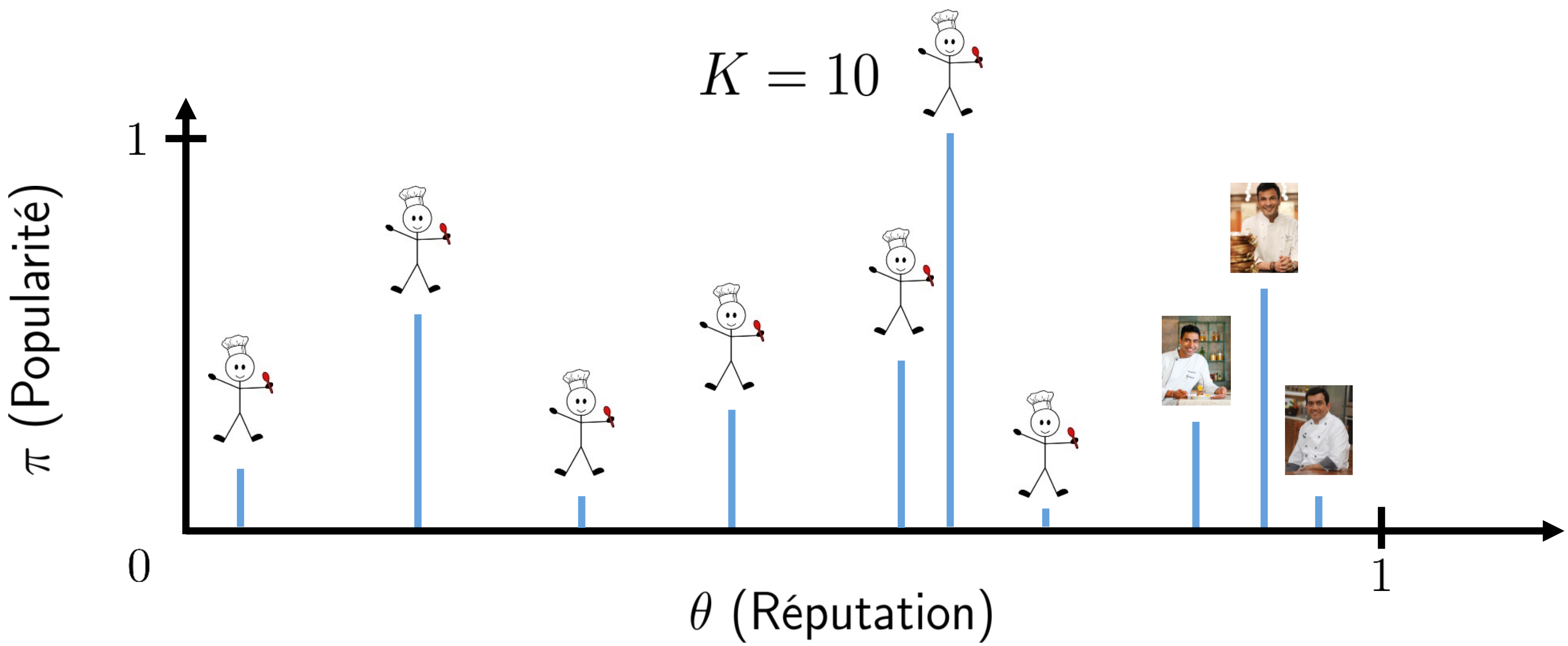
Popularité π_k

Dynamiques de connexion

$$\theta_k \sim \mathcal{U}(0,1)$$

$$\pi_k \mid \alpha, \gamma, \phi, K \sim \text{Beta}\left(\alpha\frac{\gamma}{K} + \phi \cdot \mathbb{I}(k \in O), \alpha(1 - \frac{\gamma}{K})\right)$$

$$Z_{ki} \mid \pi_k, \theta_k, \theta_i \sim \text{Bernoulli}\left(\pi_k \cdot \mathbb{I}(\theta_k > \theta_i)\right)$$

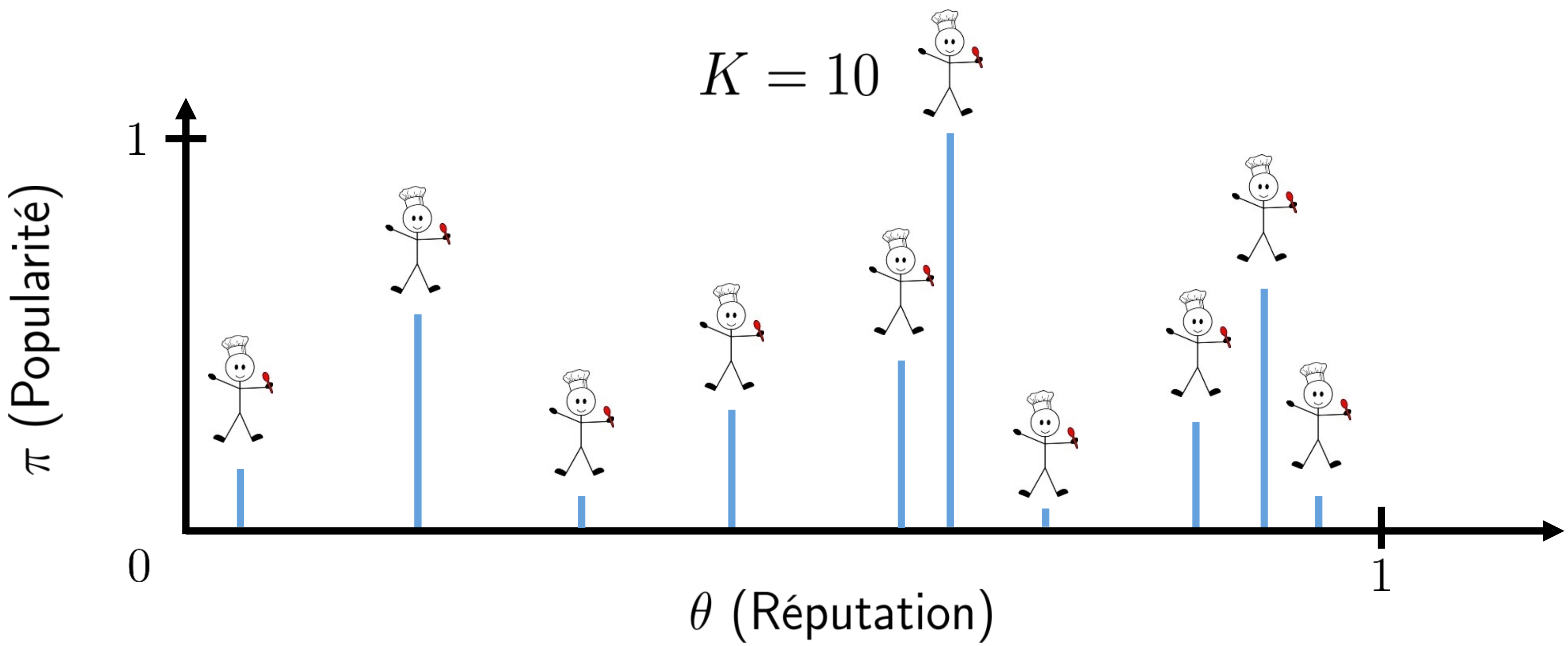


Dynamiques de connexion

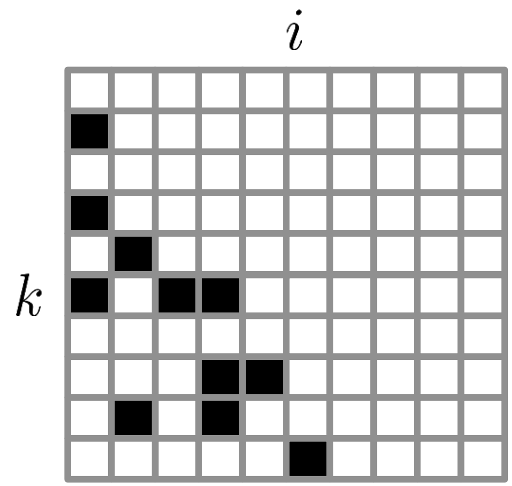
$$\theta_k \sim \mathcal{U}(0,1)$$

$$\pi_k \mid \alpha, \gamma, \phi, K \sim \text{Beta}\left(\alpha\frac{\gamma}{K} + \phi \cdot \mathbb{I}(k \in O), \alpha(1 - \frac{\gamma}{K})\right)$$

$$Z_{ki} \mid \pi_k, \theta_k, \theta_i \sim \text{Bernoulli}\left(\pi_k \cdot \mathbb{I}(\theta_k > \theta_i)\right)$$



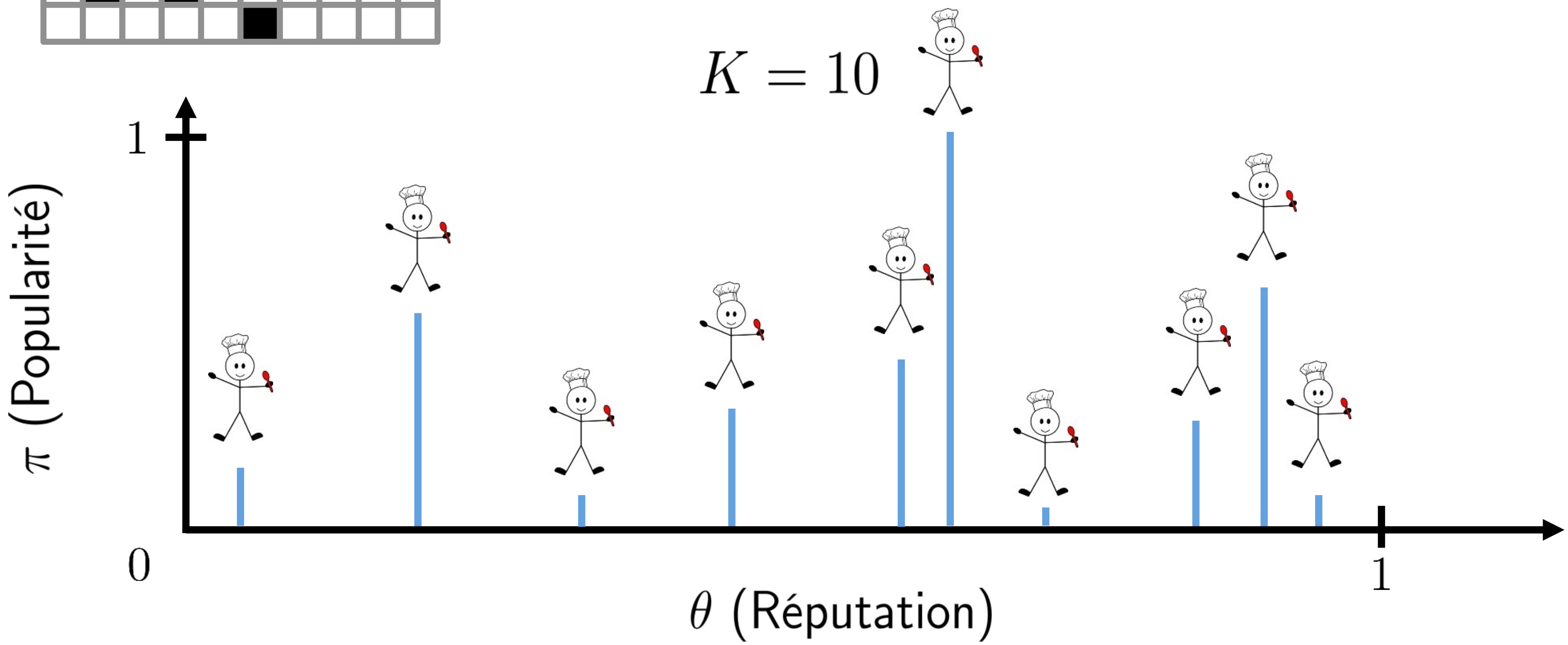
Dynamiques de connexion



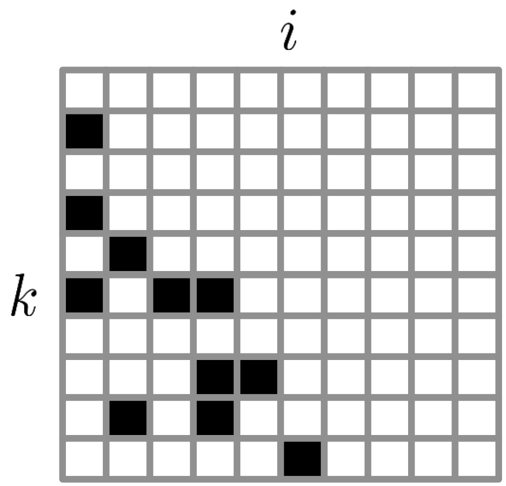
$$\theta_k \sim \mathcal{U}(0,1)$$

$$\pi_k \mid \alpha, \gamma, \phi, K \sim \text{Beta}\left(\alpha\frac{\gamma}{K} + \phi \cdot \mathbb{I}(k \in O), \alpha(1 - \frac{\gamma}{K})\right)$$

$$Z_{ki} \mid \pi_k, \theta_k, \theta_i \sim \text{Bernoulli}\left(\pi_k \cdot \mathbb{I}(\theta_k > \theta_i)\right)$$



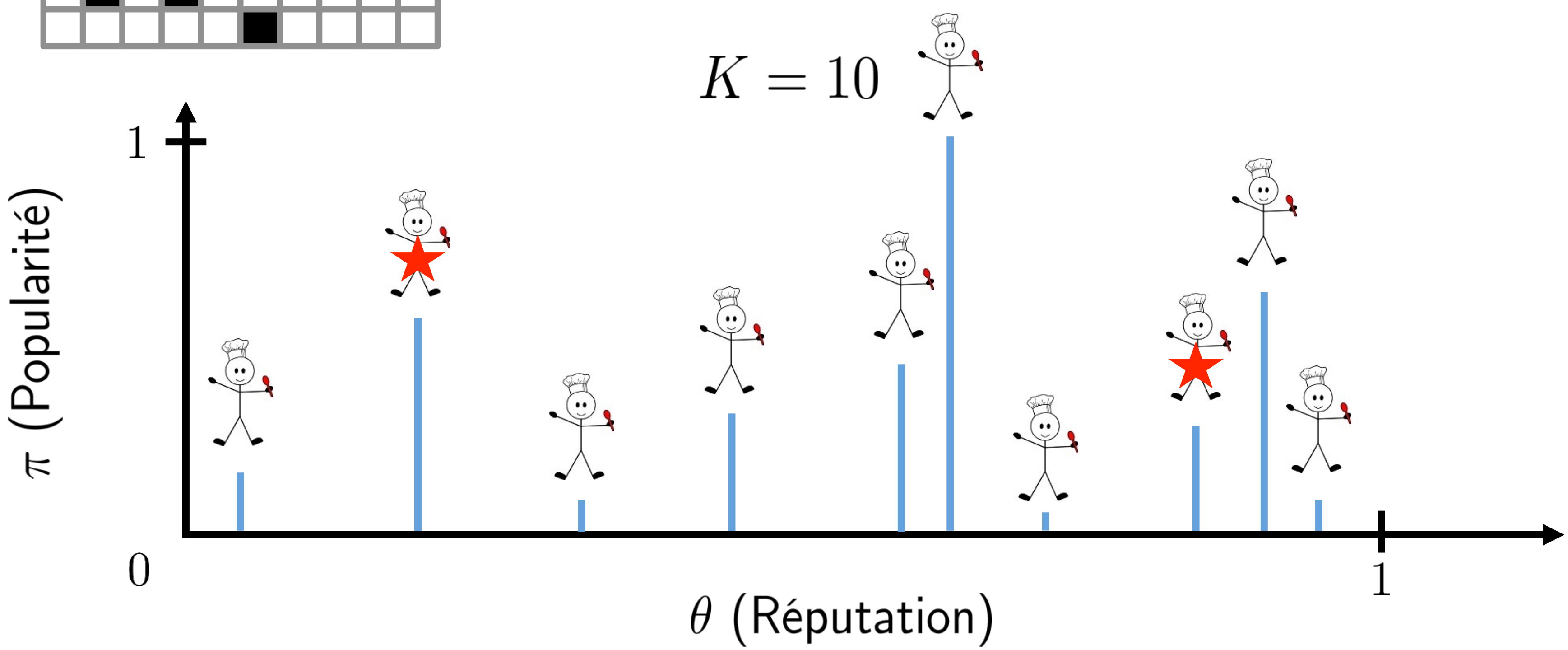
Dynamiques de connexion



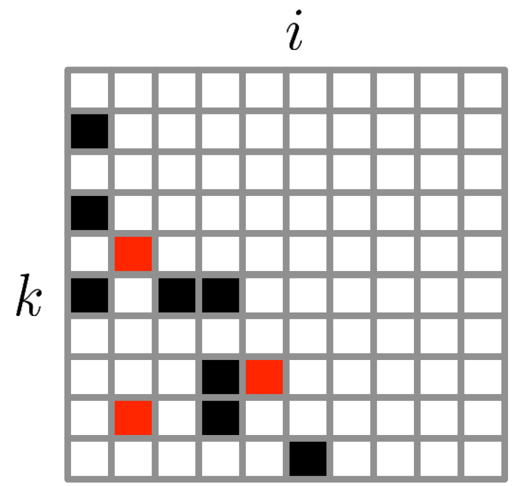
Généalogie des inspirations des chefs

$+$

Conservation des ancêtres



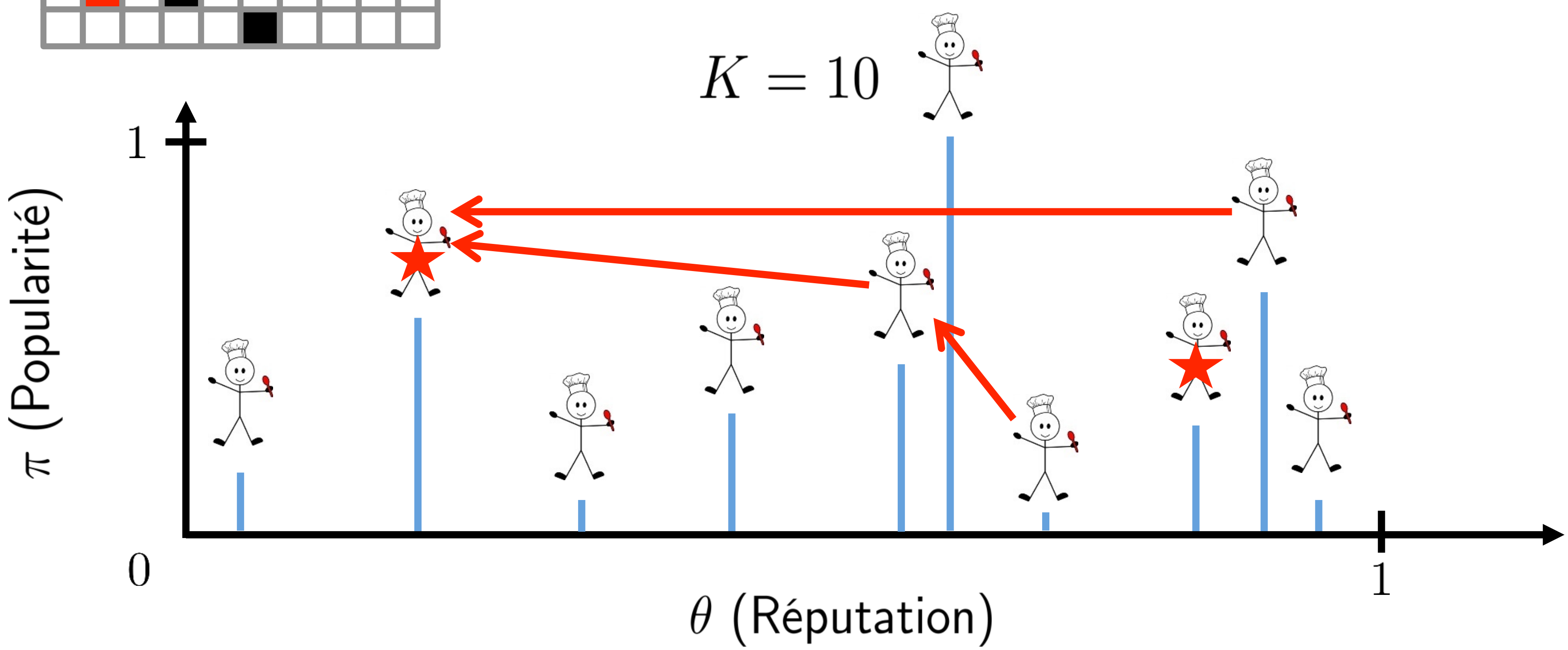
Dynamiques de connexion



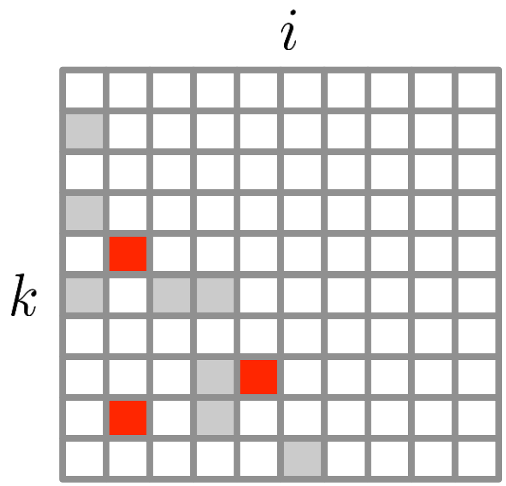
Généalogie des inspirations des chefs

$+$

Conservation des ancêtres



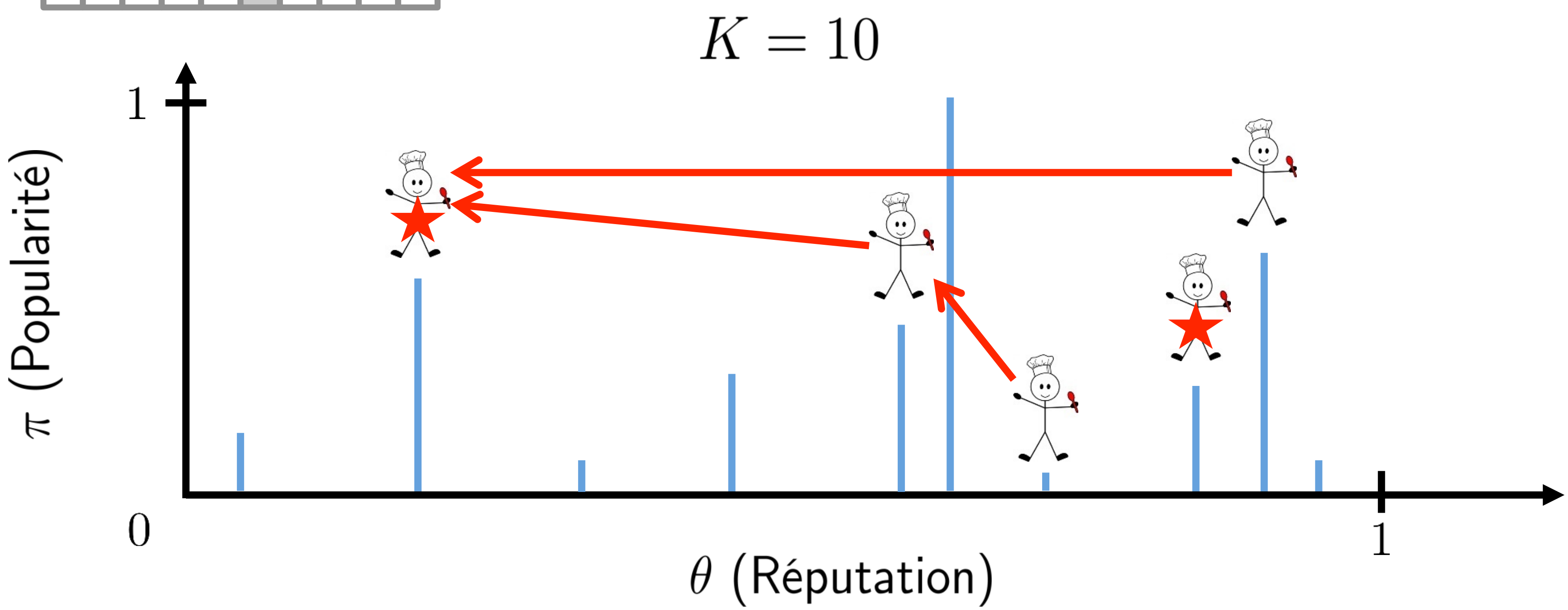
Dynamiques de connexion



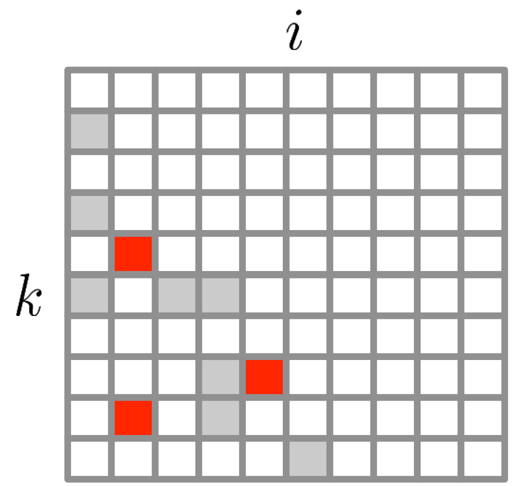
Généalogie des inspirations des chefs

+

Conservation des ancêtres



De paramétrique à nonparamétrique

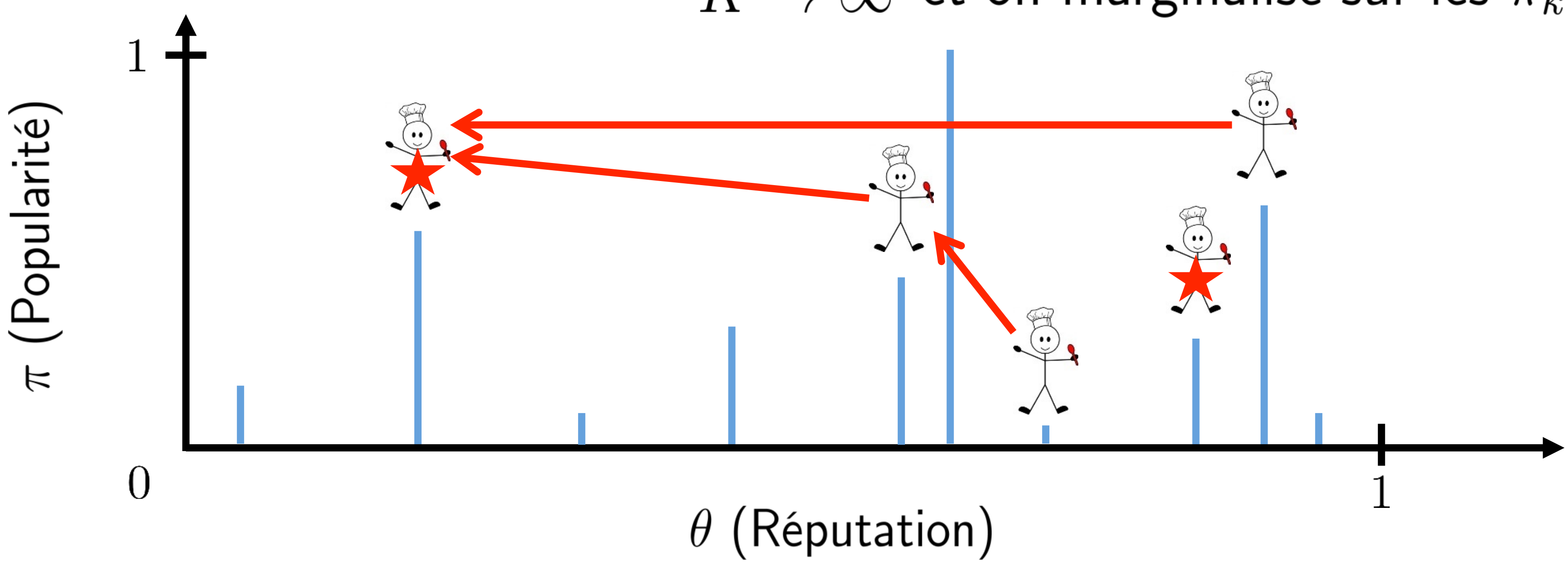


$$\theta_k \sim \mathcal{U}(0,1)$$

$$\pi_k \mid \alpha, \gamma, \phi, K \sim \text{Beta}\left(\alpha\frac{\gamma}{K} + \phi \cdot \mathbb{I}(k \in O), \alpha(1 - \frac{\gamma}{K})\right)$$

$$Z_{ki} \mid \pi_k, \theta_k, \theta_i \sim \text{Bernoulli}\left(\pi_k \cdot \mathbb{I}(\theta_k > \theta_i)\right)$$

$K \to \infty$ et on marginalise sur les π_k



Processus des chefs Indiens

- Le processus des chefs Indiens est sous-tendu par un processus Beta, une distribution sur les mesures complétement aléatoires.

- La marginalisation du processus Beta nous donne :

$$p(Z_{AA}^{\nearrow}, Z_{IA}, \boldsymbol{\theta}_A^{\nearrow} | \alpha, \gamma, \phi, O) =$$
$$\frac{1}{K^+!} \exp\left(-\alpha\gamma \sum_{j=1}^{K^+} (\theta_{j+1}^{\nearrow} - \theta_j^{\nearrow}) \left[\psi(\alpha + j) - \psi(\alpha)\right]\right)$$
$$\prod_{k \in A^+} \alpha\gamma \frac{(m_k - 1)!}{(\alpha + \downarrow_k - m_k)^{\overline{m_k}}} \prod_{k \in O} \frac{\phi^{\overline{m_k}} \alpha^{\overline{\downarrow_k - m_k}}}{[\alpha + \phi]^{\overline{\downarrow_k}}}$$

- La distribution est utilisé en tant qu'a priori sur l'espace des DAG.

- Permet de construire des opérateurs MCMC pour l'inférence

Processus des chefs Indiens

- Le processus des chefs Indiens est sous-tendu par un processus Beta, une distribution sur les mesures complétement aléatoires.

- La marginalisation du processus Beta nous donne :

réputation croissante

nombre de noeuds actifs

fonction digamma

$$p(Z^{\nearrow}_{AA}, Z_{IA}, \boldsymbol{\theta}^{\nearrow}_{A} | \alpha, \gamma, \phi, O) =$$

sous-matrice active triée

$$\frac{1}{K^+!} \exp\left(-\alpha\gamma \sum_{j=1}^{K^+} (\theta^{\nearrow}_{j+1} - \theta^{\nearrow}_{j}) \left[\psi(\alpha + j) - \psi(\alpha)\right] \right)$$

fonction factoriel croissante

enveloppe de connection nulle (inactif vers actif)

$$\prod_{k \in A^+} \alpha\gamma \frac{(m_k - 1)!}{(\alpha + \downarrow_k - m_k)^{\overline{m_k}}} \prod_{k \in O} \frac{\phi^{\overline{m_k}} \alpha^{\overline{\downarrow_k - m_k}}}{[\alpha + \phi]^{\overline{\downarrow_k}}}$$

degré sortant

noeuds cachées

noeuds observables

nombre de noeud sous k

- La distribution est utilisé en tant qu'a priori sur l'espace des DAG.

- Permet de construire des opérateurs MCMC pour l'inférence

Processus des chefs Indiens

- Le processus des chefs Indiens est sous-tendu par un processus Beta, une distribution sur les mesures complétement aléatoires.

- La marginalisation du processus Beta nous donne :

réputation croissante

nombre de noeuds actifs

fonction digamma

$$p(Z^{\nearrow}_{AA}, Z_{IA}, \boldsymbol{\theta}^{\nearrow}_{A} | \alpha, \gamma, \phi, O) =$$

sous-matrice active triée

enveloppe de connection nulle (inactif vers actif)

$$\frac{1}{K^+!} \exp\left(-\alpha\gamma \sum_{j=1}^{K^+} (\theta^{\nearrow}_{j+1} - \theta^{\nearrow}_{j}) \left[\psi(\alpha + j) - \psi(\alpha)\right] \right)$$

fonction factoriel croissante

$$\prod_{k \in A^+} \alpha\gamma \frac{(m_k - 1)!}{(\alpha + \downarrow_k - m_k)^{\overline{m_k}}} \prod_{k \in O} \frac{\phi^{\overline{m_k}} \alpha^{\overline{\downarrow_k - m_k}}}{[\alpha + \phi]^{\overline{\downarrow_k}}}$$

degré sortant

noeuds cachées

noeuds observables

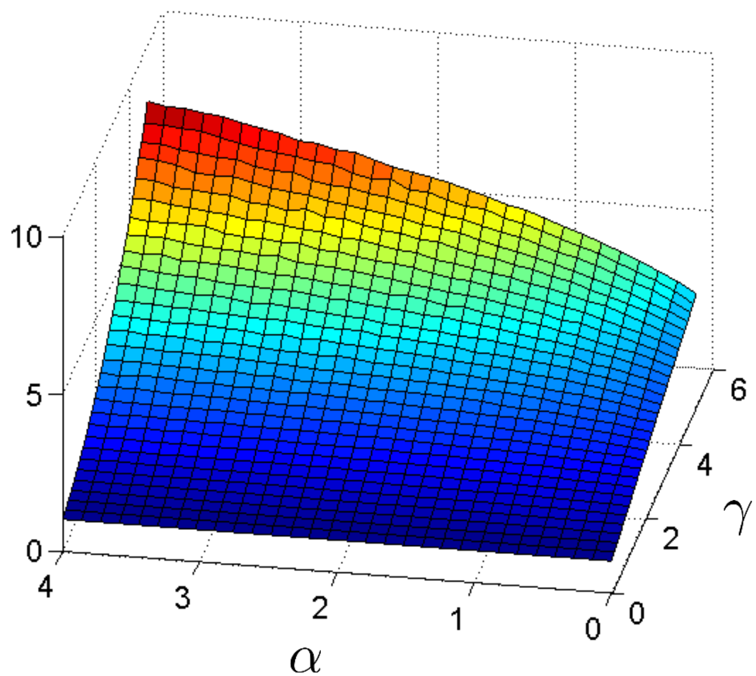
nombre de noeud sous k

- La distribution est utilisé en tant qu'a priori sur l'espace des DAG.

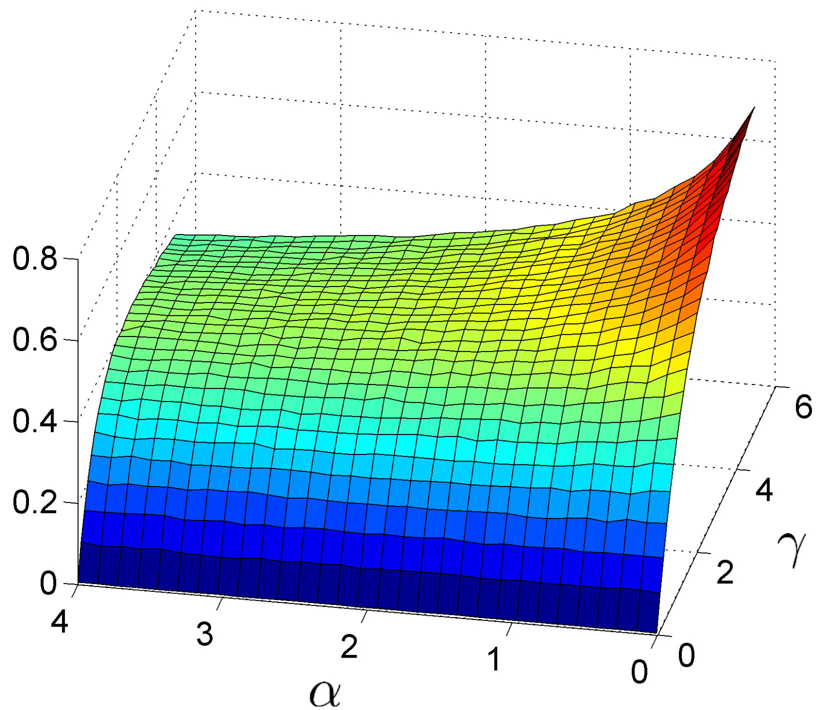
- Permet de construire des opérateurs MCMC pour l'inférence

Propriétés de la distribution

Espérence du nombre de noeuds



Nombre connections vs nombre de noeuds



Le paramètre α influence la densité de la matrice d'adjacence

Plan de présentation

- Introduction à l'apprentissage Bayesien
- Présentation du processus des chefs Indiens
- **Application à l'apprentissage de réseaux Bayésiens**
- Potentiel d'application pour apprendre la structure d'un réseau de neurones?

Apprentissage Bayésien du modèle

Définir un modèle probabiliste pour les observations :

$$p(y_i \mid x_i, \boldsymbol{\theta}, M)$$

Définir le prior sur les paramètres et les modèles :

$$p(\boldsymbol{\theta}, M) = p(\boldsymbol{\theta} \mid M)p(M)$$

Procéder à l'inférence :

$$p(\boldsymbol{\theta}, M|X, Y) = \frac{p(Y \mid \boldsymbol{\theta}, X)p(\boldsymbol{\theta} \mid M)p(M)}{\iint p(Y \mid \boldsymbol{\theta}, X)p(\boldsymbol{\theta} \mid M)p(M)\mathrm{d}\boldsymbol{\theta}\mathrm{d}M}$$

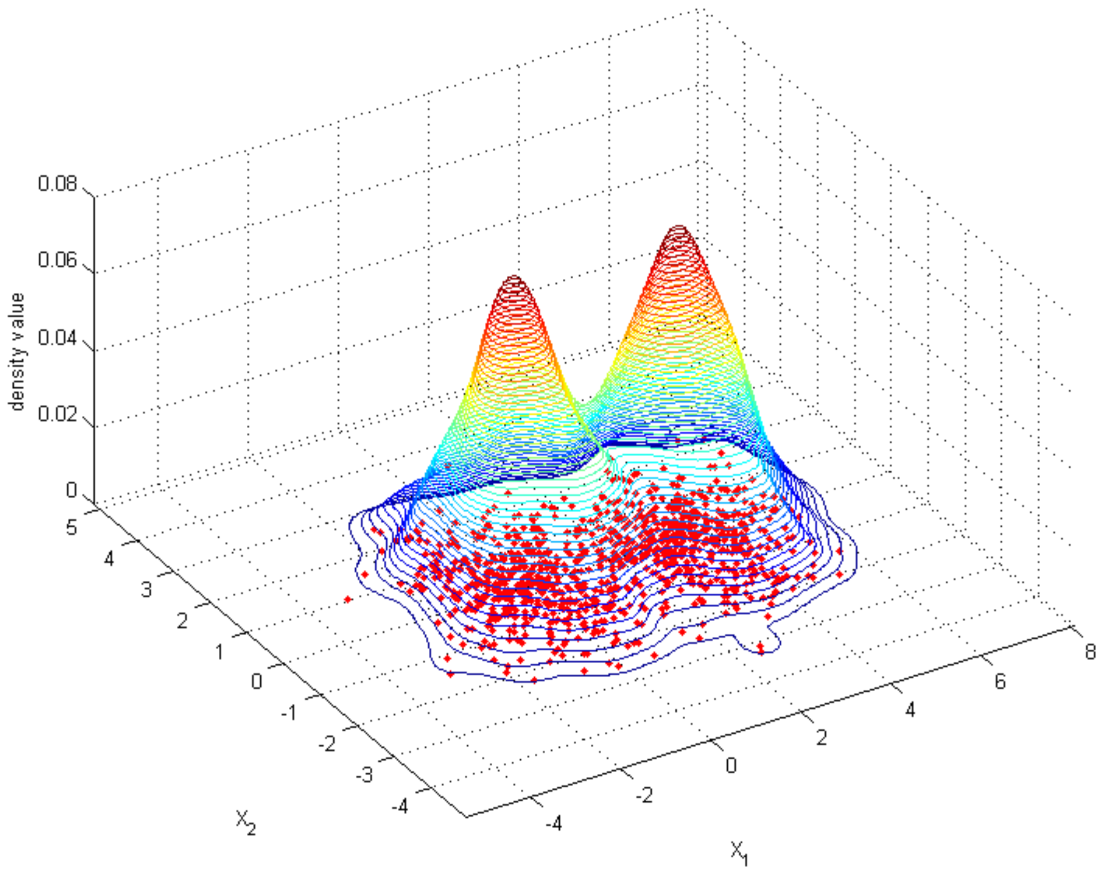
Faire une prédiction à partir de l'*a posteriori* :

$$p(y_* \mid x_*, X, Y) = \iint p(y_* \mid x_*, \boldsymbol{\theta}, M)p(\boldsymbol{\theta}, M \mid X, Y)\mathrm{d}\boldsymbol{\theta}$$

Estimation de densité

Objectif : modéliser la probabilité jointe d'un ensemble de variable

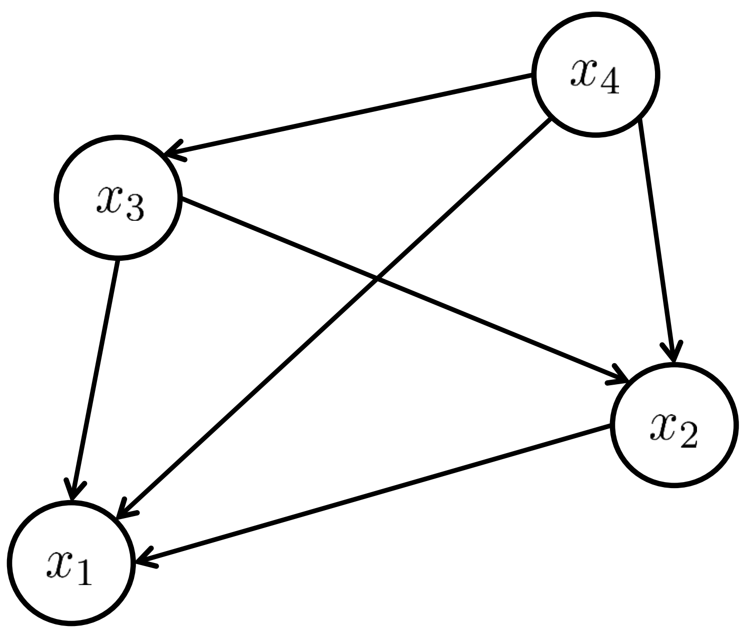
$$p(\mathbf{x})$$



Réseaux Bayésien

Un réseaux Bayésien est une représentation graphique d'une factorisation d'une distribution de probabilité jointe

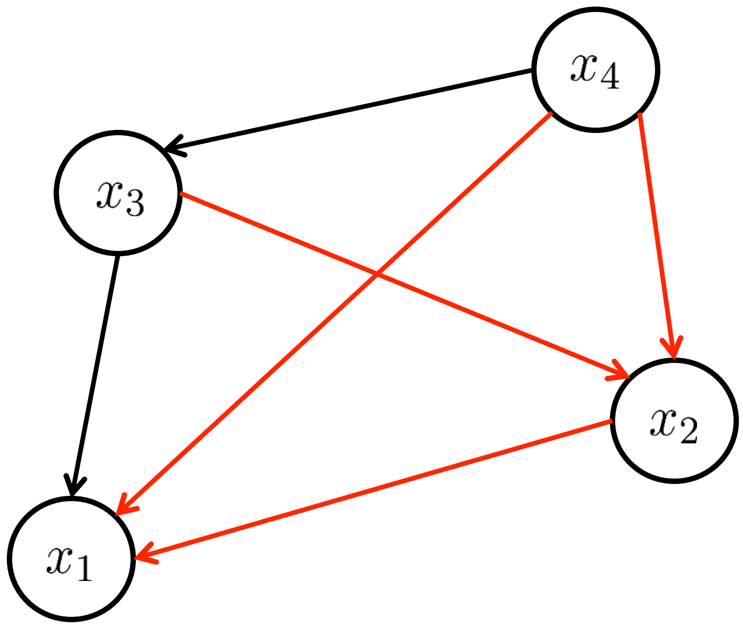
$$p(x_1, x_2, x_3, x_4) = p(x_1 \mid x_2, x_3, x_4)p(x_2 \mid x_3, x_4)p(x_3 \mid x_4)p(x_4)$$



Réseaux Bayésien

Un réseaux Bayésien est une représentation graphique d'une factorisation d'une distribution de probabilité jointe

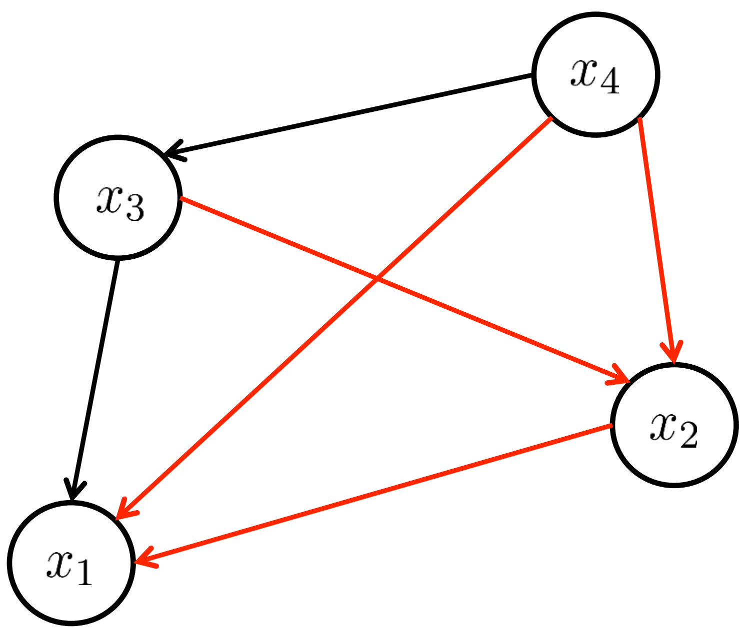
$$p(x_1, x_2, x_3, x_4) = p(x_1 \mid x_2, x_3, x_4)p(x_2 \mid x_3, x_4)p(x_3 \mid x_4)p(x_4)$$



Réseaux Bayésien

Un réseaux Bayésien est une représentation graphique d'une factorisation d'une distribution de probabilité jointe

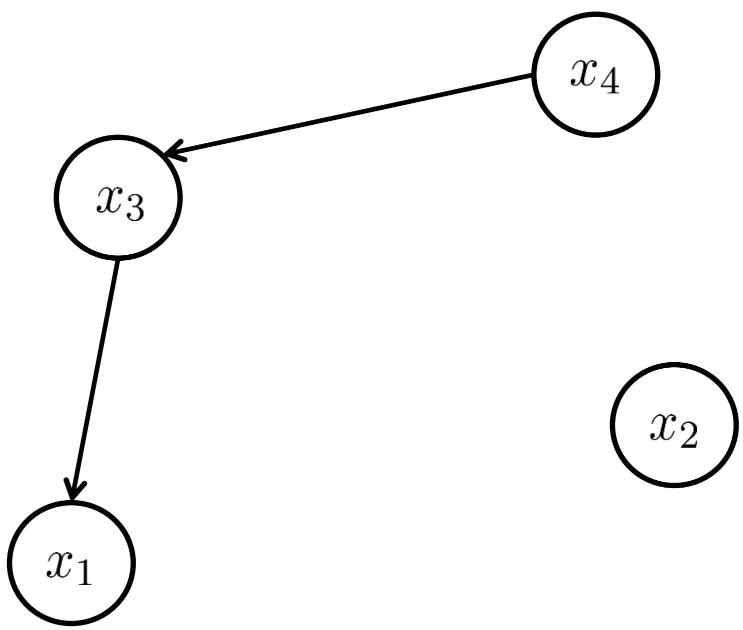
$$p(x_1, x_2, x_3, x_4) = p(x_1 \mid x_3)p(x_2)p(x_3 \mid x_4)p(x_4)$$



Réseaux Bayésien

Un réseaux Bayésien est une représentation graphique d'une factorisation d'une distribution de probabilité jointe

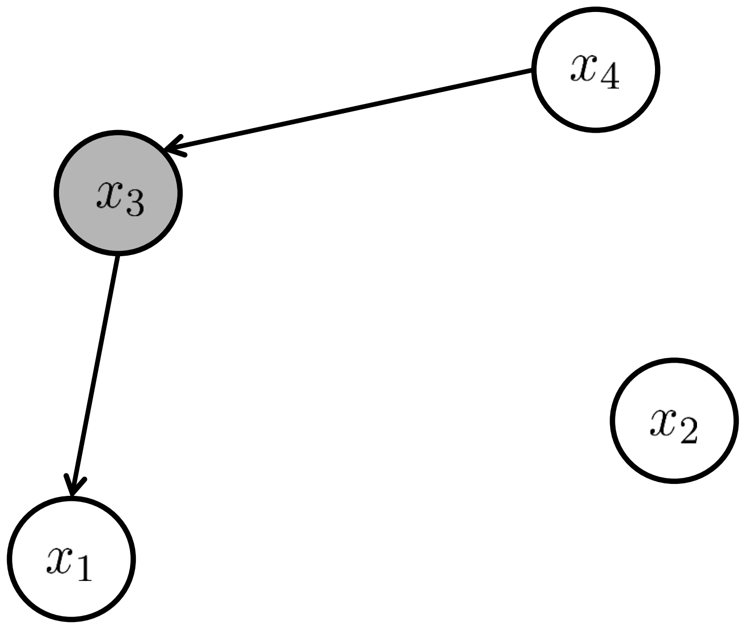
$$p(x_1, x_2, x_3, x_4) = p(x_1 \mid x_3)p(x_2)p(x_3 \mid x_4)p(x_4)$$



Réseaux Bayésien

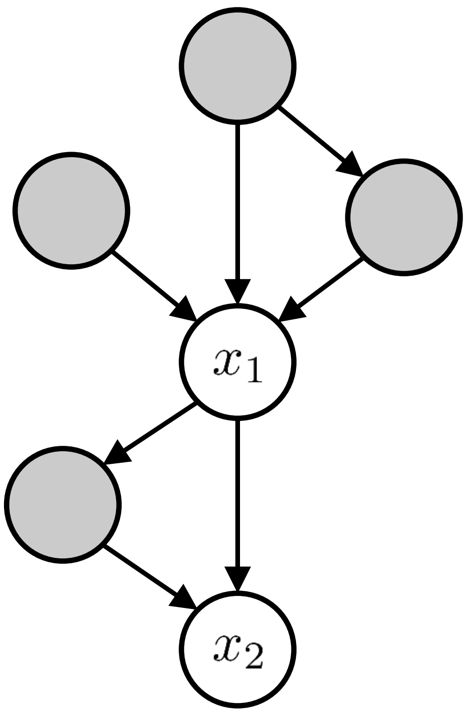
Un réseaux Bayésien est une représentation graphique d'une factorisation d'une distribution de probabilité jointe

$$p(x_1, x_2, x_3, x_4) = p(x_1 \mid x_3)p(x_2)p(x_3 \mid x_4)p(x_4)$$



Les variables peuvent être observables ou cachées

Nonlinear Gaussian Belief Networks

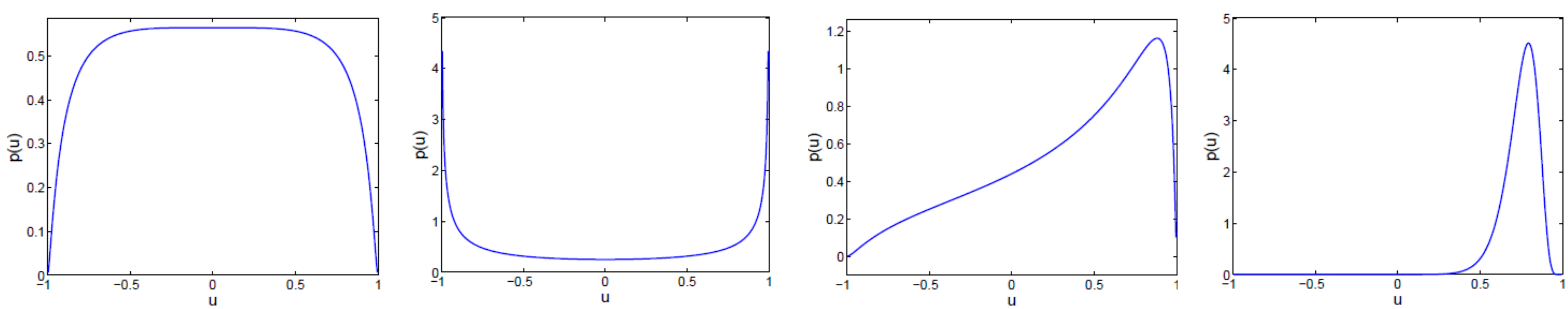


Combinaison linéaire de la valeur des parents:

$$s_i = b_i - \sum_j Z_{ji} W_{ji} x_j$$

Transformation nonlinéaire d'une Gaussienne $\mathcal{N}(s_i, 1/\rho_i)$ dans une sigmoide :

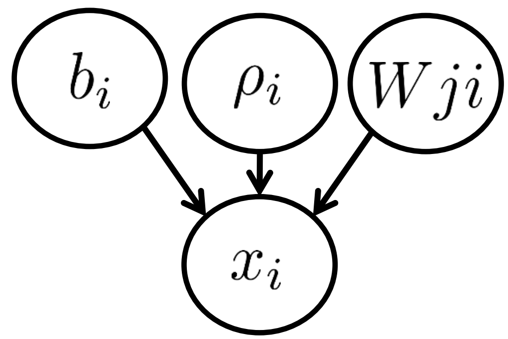
$$p(x_i \mid s_i, \rho_i) = \frac{\exp\left\{-\frac{\rho_i}{2}\left[\sigma^{-1}(x_i) - s_i\right]^2\right\}}{\sigma'(\sigma^{-1}(x_i))\sqrt{\frac{2\pi}{\rho_i}}},$$



Apprentissage Bayesien du modèle

Définir un modèle probabiliste pour les observations :

$$p(y_i \mid x_i, \boldsymbol{\theta}, M)$$



Définir le prior sur les paramètres et les modèles :

$$p(\boldsymbol{\theta}, M) = p(\boldsymbol{\theta} \mid M)p(M)$$

Procéder à l'inférence :

$$p(\boldsymbol{\theta}, M|X, Y) = \frac{p(Y \mid \boldsymbol{\theta}, X)p(\boldsymbol{\theta} \mid M)p(M)}{\iint p(Y \mid \boldsymbol{\theta}, X)p(\boldsymbol{\theta} \mid M)p(M)\mathrm{d}\boldsymbol{\theta}\mathrm{d}M}$$

Faire une prédiction à partir de l'*a posteriori* :

$$p(y_* \mid x_*, X, Y) = \iint p(y_* \mid x_*, \boldsymbol{\theta}, M)p(\boldsymbol{\theta}, M \mid X, Y)\mathrm{d}\boldsymbol{\theta}$$

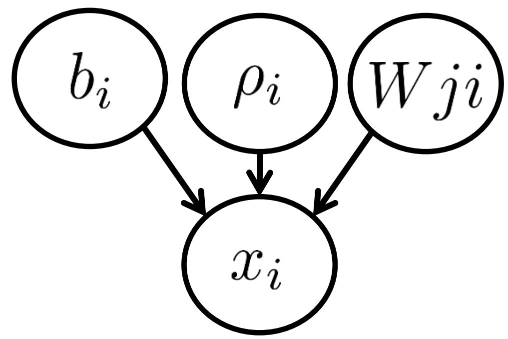
Apprentissage Bayésien du modèle

Définir un modèle probabiliste pour les observations :

$$p(y_i \mid x_i, \boldsymbol{\theta}, M)$$

Définir le prior sur les paramètres et les modèles :

$$p(\boldsymbol{\theta}, M) = p(\boldsymbol{\theta} \mid M)p(M)$$



Procéder à l'inférence :

$$\text{MCMC} \quad p(\boldsymbol{\theta}, M|X, Y) = \frac{p(Y \mid \boldsymbol{\theta}, X)p(\boldsymbol{\theta} \mid M)p(M)}{\iint p(Y \mid \boldsymbol{\theta}, X)p(\boldsymbol{\theta} \mid M)p(M)\mathrm{d}\boldsymbol{\theta}\mathrm{d}M}$$

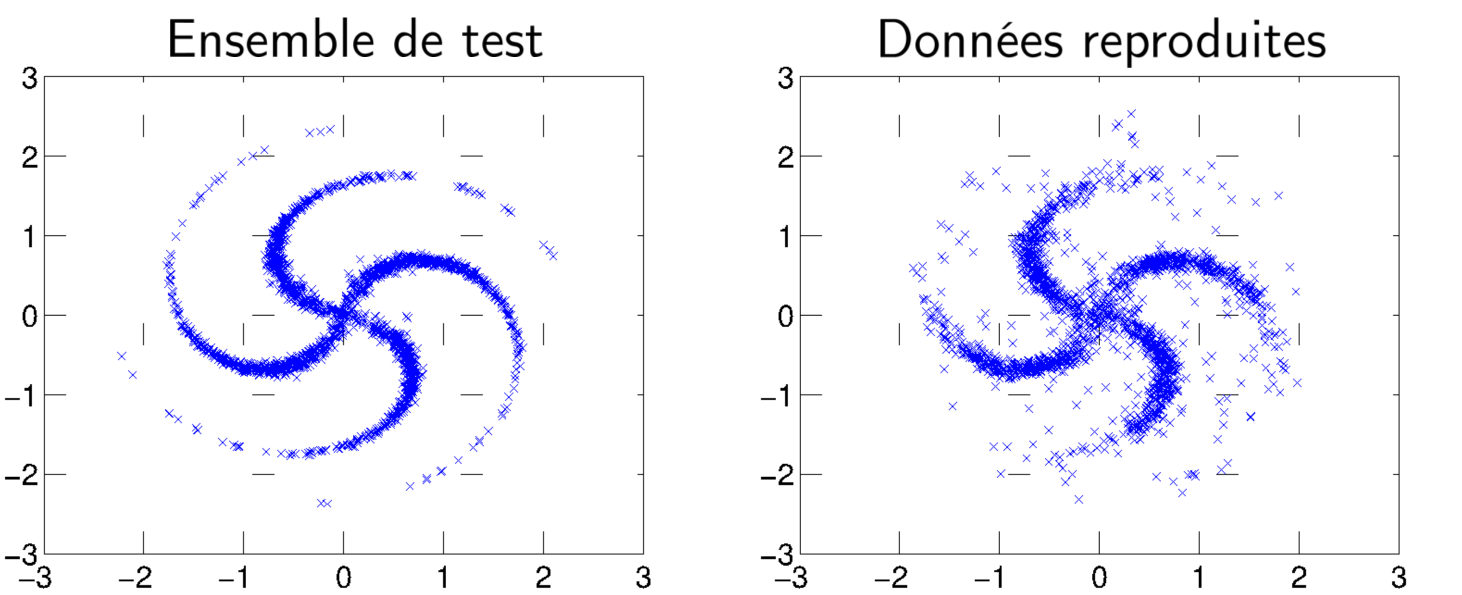
Faire une prédiction à partir de l'*a posteriori* :

$$p(y_* \mid x_*, X, Y) = \iint p(y_* \mid x_*, \boldsymbol{\theta}, M)p(\boldsymbol{\theta}, M \mid X, Y)\mathrm{d}\boldsymbol{\theta}$$

Résultats sur l'estimation de densité

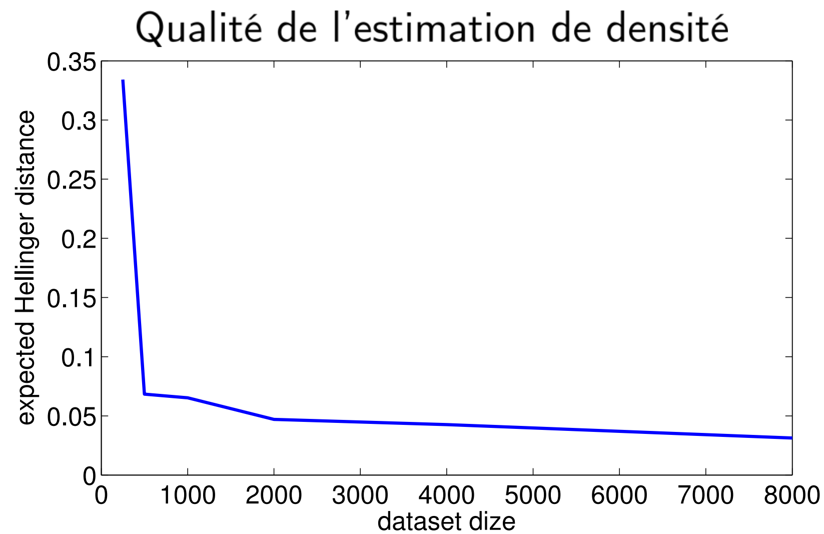
Table 1: Estimation de la distance de Hellinger entre ensembles de test et fantaisie.

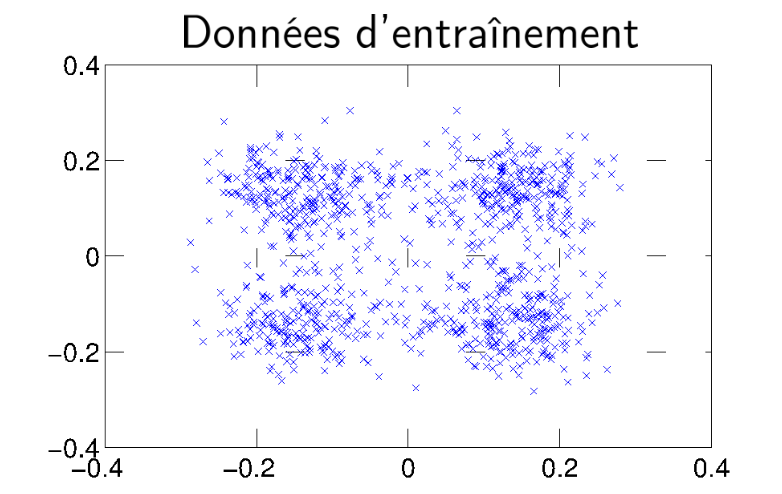
	Test sets				
	Ring	Two Moons	Pinwheel	Geyser	Iris
ICP	**0.0402**	**0.0342**	**0.0547**	**0.0734**	0.2666
CIBP	0.0493	0.0469	0.0692	0.1246	0.2667
ECIBP	0.0419	0.0450	0.0685	0.1171	**0.2632**
Training set	0.0312	0.0138	0.0436	0.0234	0.1930

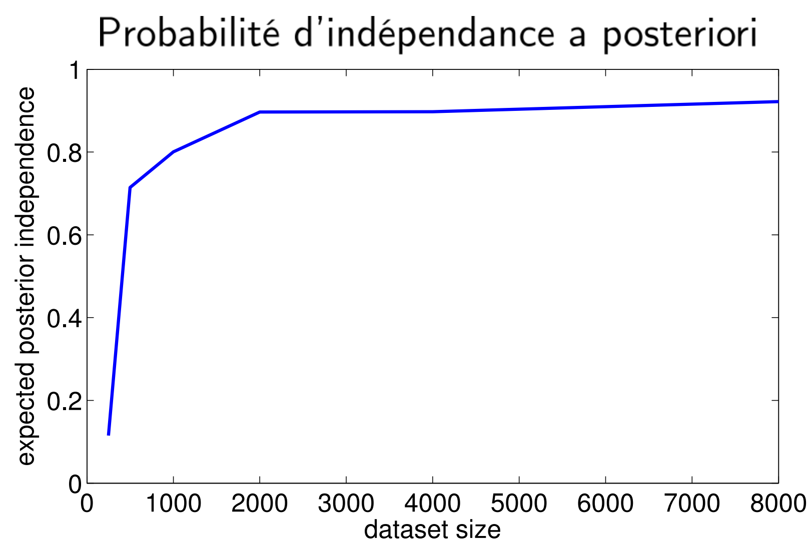


Résultats de tests d'indépendances

- Distribution quadrimodale sur 2 dimensions indépendantes
- Échantillonnage de structures a posteriori
- Test d'indépendance avec la d-séparation
- Calculs d'espérances sur l'indépendance





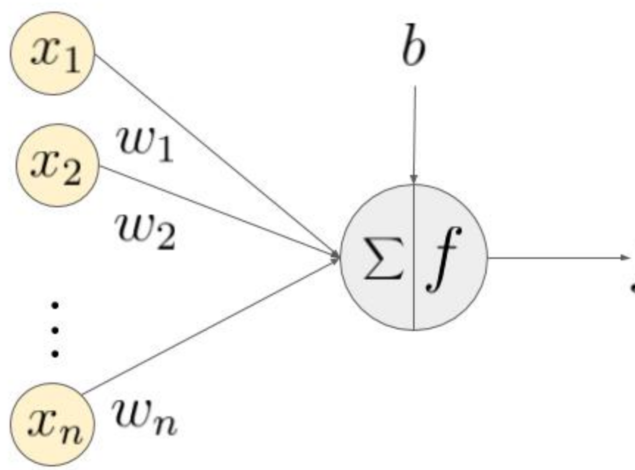


Plan de présentation

- Introduction à l'apprentissage Bayesien
- Présentation du processus des chefs Indiens
- Application à l'apprentissage de réseaux Bayésiens
- **Potentiel d'application pour apprendre la structure d'un réseau de neurones?**

Paramétrisation d'un réseau de neurones

Des poids

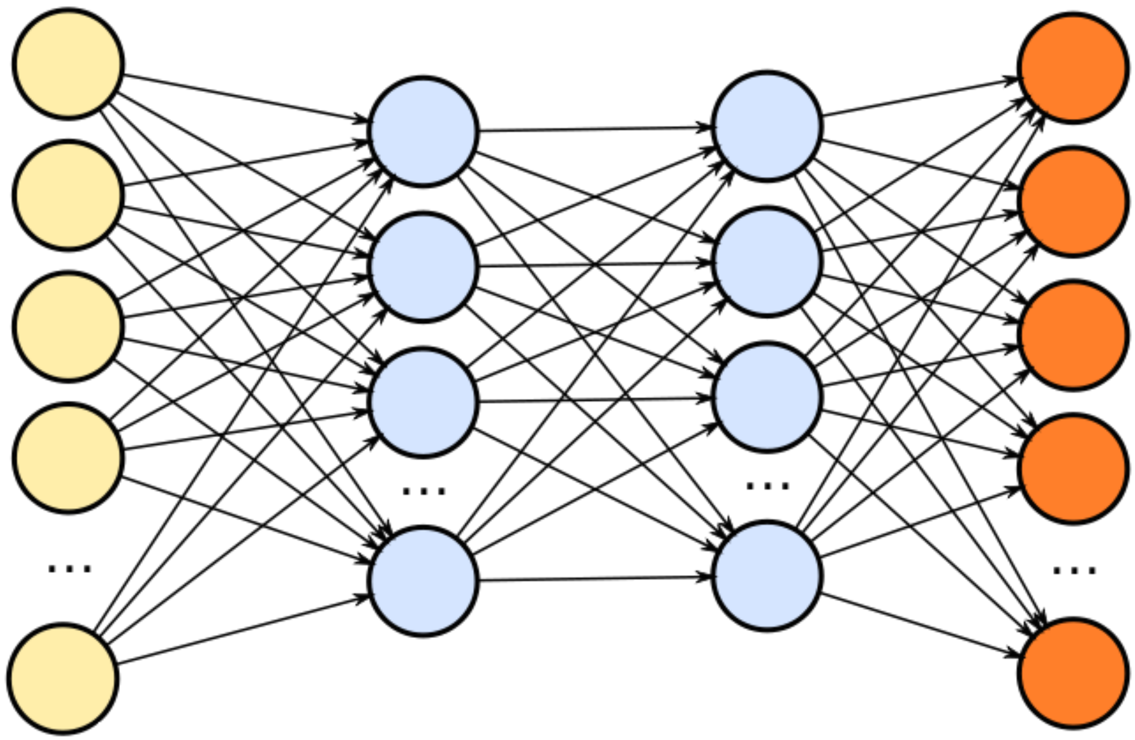


$$f\left(b + \sum_{i=1}^{n} x_i w_i\right)$$

Des biais

Structure d’un réseau de neurones

Un graphe



wp.wwu.edu

Apprentissage Bayesien du modèle

Définir un modèle probabiliste pour les observations :

$$p(y_i \mid x_i, \boldsymbol{\theta}, M)$$

réseaux de neurones avec probabilité en sortie

Définir le prior sur les paramètres et les modèles :

$$p(\boldsymbol{\theta}, M) = p(\boldsymbol{\theta} \mid M) p(M)$$

Processus des chefs Indiens

Des gaussiennes

Procéder à l'inférence :

$$p(\boldsymbol{\theta}, M | X, Y) = \frac{p(Y \mid \boldsymbol{\theta}, X) p(\boldsymbol{\theta} \mid M) p(M)}{\iint p(Y \mid \boldsymbol{\theta}, X) p(\boldsymbol{\theta} \mid M) p(M) \mathrm{d}\boldsymbol{\theta} \mathrm{d}M}$$

Faire une prédiction à partir de l'*a posteriori* :

$$p(y_* \mid x_*, X, Y) = \iint p(y_* \mid x_*, \boldsymbol{\theta}, M) p(\boldsymbol{\theta}, M \mid X, Y) \mathrm{d}\boldsymbol{\theta}$$

Vers une inférence plus rapide…

$$p(\boldsymbol{\theta}, M \mid X, Y) \propto \prod_{i=1}^{N} p(y_i \mid \boldsymbol{\theta}, X) p(\boldsymbol{\theta} \mid M) p(M)$$

$$p(\boldsymbol{\theta}, M \mid X, Y) \propto \prod_{i \in A} p(y_i \mid \boldsymbol{\theta}, X_A) \underbrace{\overbrace{\boxed{\prod_{i \in B} p(y_i \mid \boldsymbol{\theta}, X_B) p(\boldsymbol{\theta} \mid M)}}^{\text{a posteriori intermédiaire}}}_{p(\boldsymbol{\theta} \mid M, X_B, Y_B)} p(M)$$

1. Le posterior intermédiaire est notre nouveau prior sur $\boldsymbol{\theta}$
2. Le *support* du nouveau prior doit être le même que celui du prior original
3. Une approximation $q(\boldsymbol{\theta} \mid M, X_B, Y_B)$ est acceptable (c'est un prior!)
4. La forme analytique n'est pas obligatoire pour $q(\boldsymbol{\theta} \mid M, X_B, Y_B)$
5. L'échantillonnage de $q(\boldsymbol{\theta} \mid M, X_B, Y_B)$ doit être **rapide**

Considération sur le pseudo-prior

$$p(\boldsymbol{\theta}, M \mid X, Y) \propto \prod_{i \in A} p(y_i \mid \boldsymbol{\theta}, X_A) \boxed{q(\boldsymbol{\theta} \mid M, X_B, Y_B)} p(M)$$

$$q(\boldsymbol{\theta} \mid M, X_B, Y_B) = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid M, X_B, Y_B)$$

1. Le posterior intermédiaire est notre nouveau prior sur $\boldsymbol{\theta}$
2. Le *support* du nouveau prior doit être le même que celui du prior original
3. Une approximation $q(\boldsymbol{\theta} \mid M, X_B, Y_B)$ est acceptable (c'est un prior!)
4. La forme analytique n'est pas obligatoire pour $q(\boldsymbol{\theta} \mid M, X_B, Y_B)$
5. L'échantillonnage de $q(\boldsymbol{\theta} \mid M, X_B, Y_B)$ doit être **rapide**

Considération sur le pseudo-prior

$$p(\boldsymbol{\theta}, M \mid X, Y) \propto \prod_{i \in A} p(y_i \mid \boldsymbol{\theta}, X_A) \boxed{q(\boldsymbol{\theta} \mid M, X_B, Y_B)} p(M)$$

$$\begin{aligned} \boldsymbol{\theta}_0 &\sim p(\boldsymbol{\theta} \mid M) \\ \boldsymbol{\theta}_{\text{opt}} &\sim \text{SGD}(\boldsymbol{\theta}_0, M, X_B, Y_B) \\ \widetilde{\boldsymbol{\theta}} &\sim \text{MCMC}(\boldsymbol{\theta}_{\text{opt}}, M, X_B, Y_B, T) \end{aligned}$$

1. Le posterior intermédiaire est notre nouveau prior sur $\boldsymbol{\theta}$
2. Le *support* du nouveau prior doit être le même que celui du prior original
3. Une approximation $q(\boldsymbol{\theta} \mid M, X_B, Y_B)$ est acceptable (c'est un prior!)
4. La forme analytique n'est pas obligatoire pour $q(\boldsymbol{\theta} \mid M, X_B, Y_B)$
5. L'échantillonnage de $q(\boldsymbol{\theta} \mid M, X_B, Y_B)$ doit être **rapide**

Considération sur le pseudo-prior

$$p(\boldsymbol{\theta}, M \mid X, Y) \propto \prod_{i \in A} p(y_i \mid \boldsymbol{\theta}, X_A) \boxed{q(\boldsymbol{\theta} \mid M, X_B, Y_B)} p(M)$$

$$\begin{aligned}
\boldsymbol{\theta}_0 &\sim p(\boldsymbol{\theta} \mid M) \\
\boldsymbol{\theta}_{\mathsf{opt}} &\sim \mathsf{SGD}(\boldsymbol{\theta}_0, M, X_B, Y_B) \\
\widetilde{\boldsymbol{\theta}} &\sim \mathsf{MCMC}(\boldsymbol{\theta}_{\mathsf{opt}}, M, X_B, Y_B, T)
\end{aligned}$$

« we show how to adjust the tuning parameters of constant SGD to best match the stationary distribution to a posterior, minimizing the Kullback-Leibler divergence between these two distributions. »

Mandt, S., Hoffman, M. D., and Blei, D. M. (2017). Stochastic Gradient Descent as Approximate Bayesian Inference. Journal of Machine Learning Research

Inférence de la structure

Objectif : $p(M \mid X, Y) = \int \prod_{i \in A} p(y_i \mid \boldsymbol{\theta}, X_A) q(\boldsymbol{\theta} \mid M, X_B, Y_B) p(M) \mathrm{d}\boldsymbol{\theta}$

1) Initialiser aléatoirement $\widetilde{M}_0$ et fixer à $t = 0$

2) Générer un candidat M' à partir du *proposal* $g(M' \mid \widetilde{M}_t)$

3) Calculer la probabilité d'acceptation :

(nécessite une estimation non-biaisé du posterior)

$$A(M' \mid \widetilde{M}_t) = \min\left(1, \frac{p(M' \mid X, Y) g(\widetilde{M}_t \mid M')}{p(\widetilde{M} \mid X, Y)\, g(M' \mid \widetilde{M}_t)}\right)$$

4) Avec probabilité $A(M' \mid \widetilde{M}_t)$, accepter et $\widetilde{M}_{t+1} = M'$, sinon rejeter et $\widetilde{M}_{t+1} = \widetilde{M}_t$

Conclusion

$$p(\boldsymbol{\theta}, M \mid X, Y) \propto \prod_{i \in A} p(y_i \mid \boldsymbol{\theta}, X_A) \boxed{q(\boldsymbol{\theta} \mid M, X_B, Y_B)} p(M)$$

$$\begin{aligned}
\boldsymbol{\theta}_0 &\sim p(\boldsymbol{\theta} \mid M) \\
\boldsymbol{\theta}_{\text{opt}} &\sim \text{SGD}(\boldsymbol{\theta}_0, M, X_B, Y_B) \\
\widetilde{\boldsymbol{\theta}} &\sim \text{MCMC}(\boldsymbol{\theta}_{\text{opt}}, M, X_B, Y_B, T)
\end{aligned}$$

1) Pour T grand, l'estimation du posterior intermédiaire est non biaisé

2) Si SGD est proche de la distribution stationnaire, alors T peut être petit

Est-ce qu'on tient un algorithme d'inférence de structures rapide efficace?

Infinite Neural Network?

Machine Learning

L'approche machine learning standard:

1. Définir un modèle prédictif $\hat{y} = f_\theta(x)$

2. Choisir une fonction de coût $L(\hat{y}, y)$

3. Trouver f_θ qui minimise le risque empirique

Exemple - Régression Linéaire

L'approche machine learning standard:

1. Définir un modèle prédictif $\hat{y} = f_\theta(x)$

$$\hat{y} = \theta_1 + \theta_2 x \qquad \text{(modèle linéaire)}$$

2. Choisir une fonction de coût $L(\hat{y}, y)$

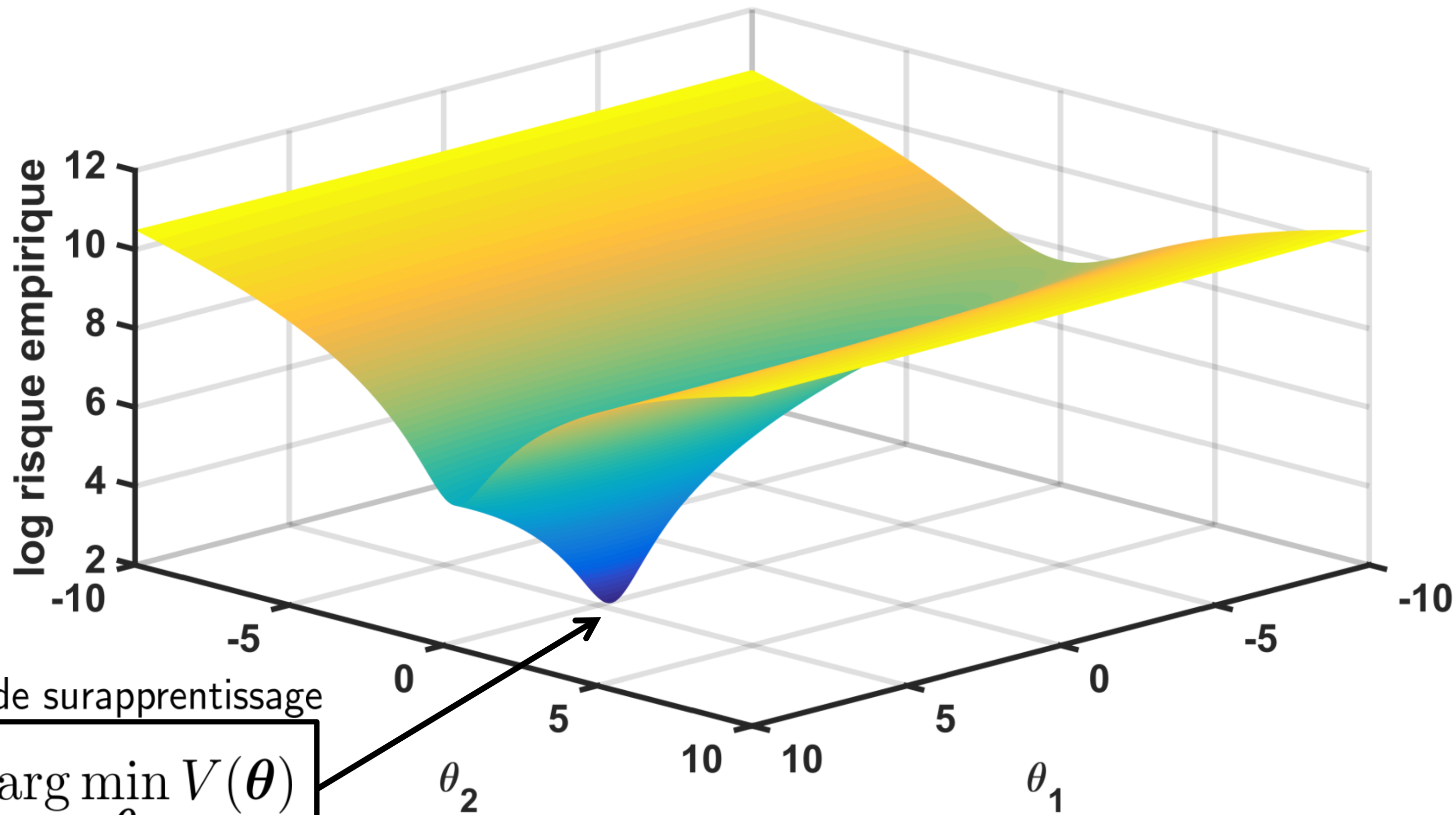
$$L(\hat{y}, y) = (\hat{y} - y)^2 \qquad \text{(erreur quadratique)}$$

3. Trouver f_θ qui minimise le risque empirique

$$V(\boldsymbol{\theta}) = \frac{1}{N} \sum_{n=1}^{N} L(\hat{y}_n, y_n)$$

Fonction de risque empirique

$$V(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N}(f_\theta(x_n) - y_n)^2$$



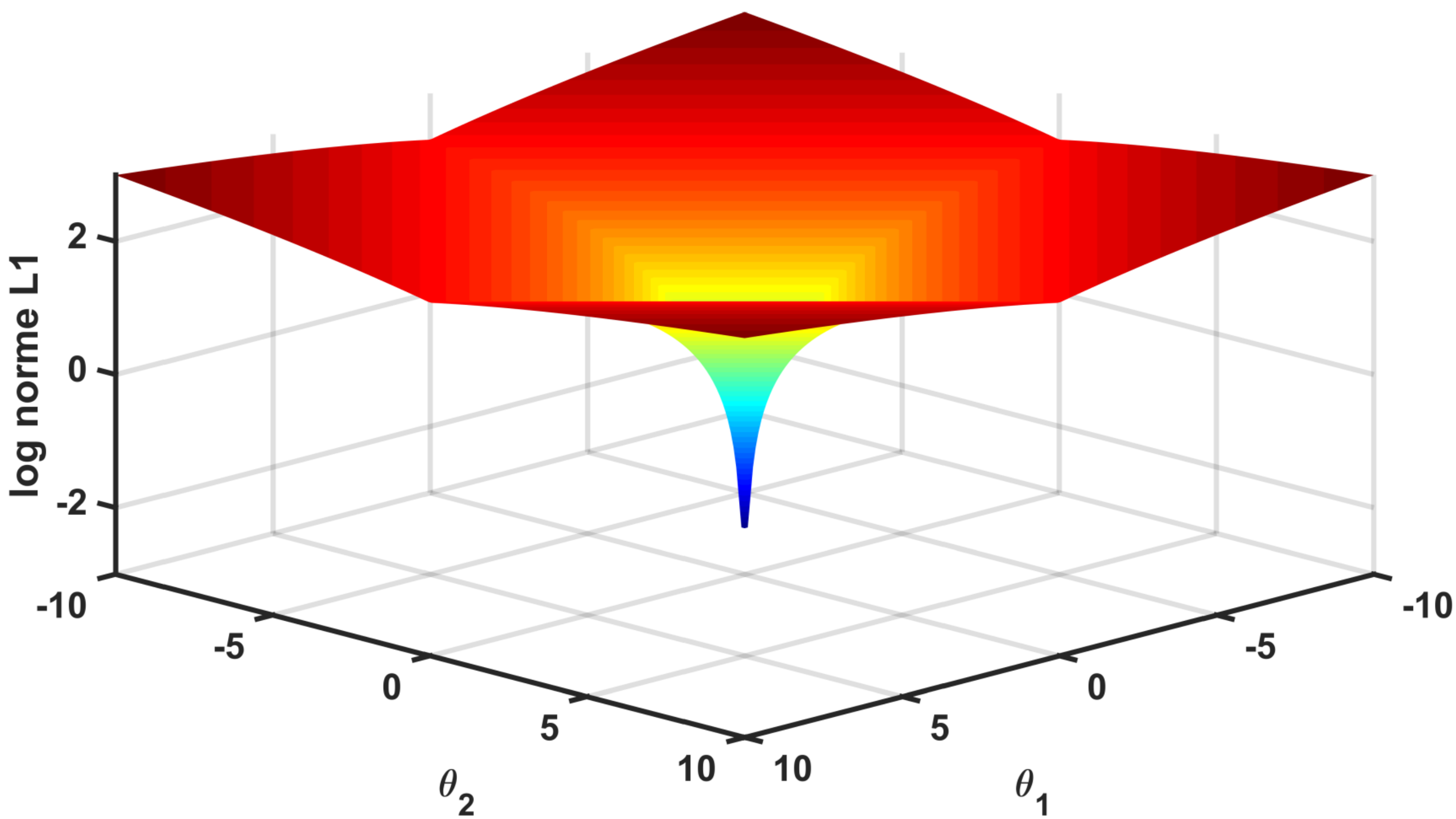
Risque de surapprentissage

$$\boxed{\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} V(\boldsymbol{\theta})}$$

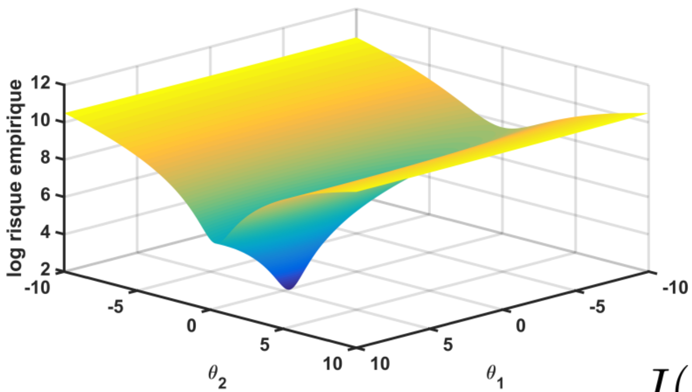
Régularisation

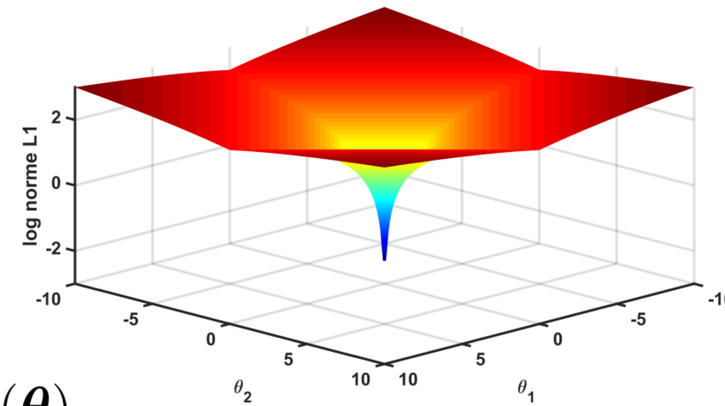
La régularisation aide à limiter le surapprentissage

régularisation L1 : $R(\boldsymbol{\theta}) = \sum_j |\theta_j|$

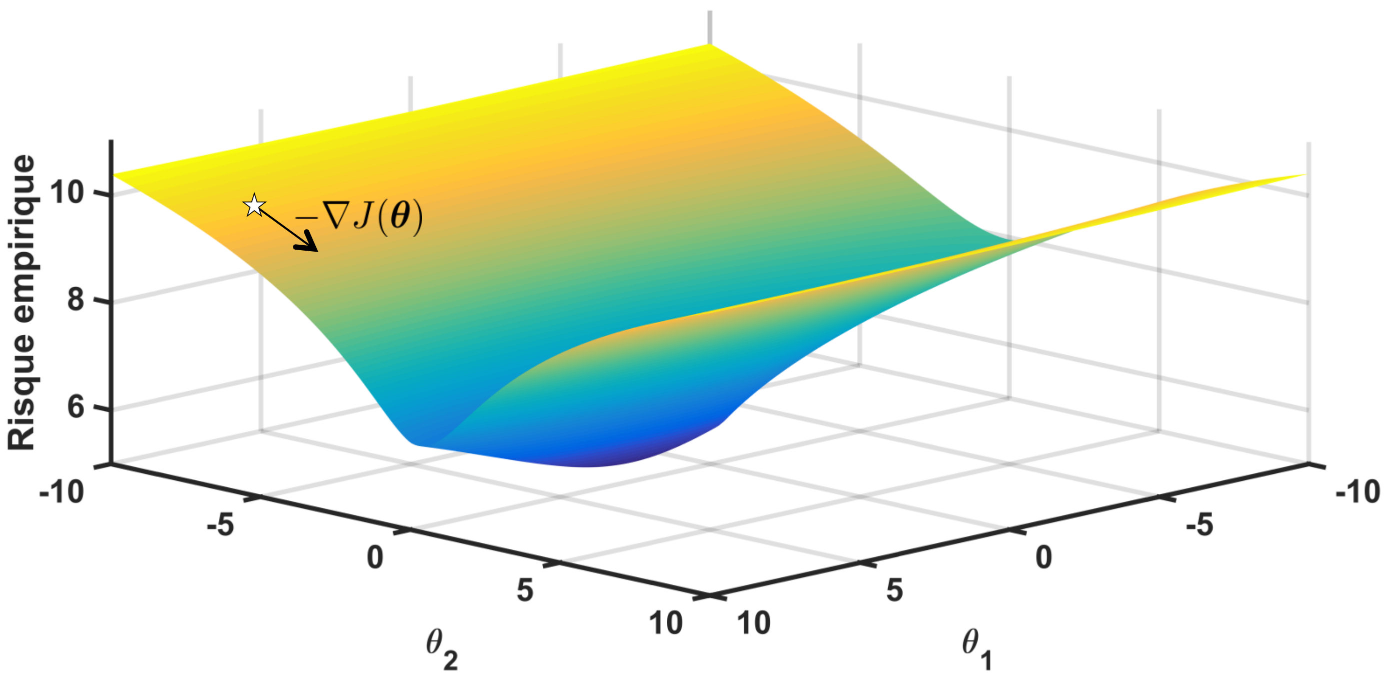


La fonction objective



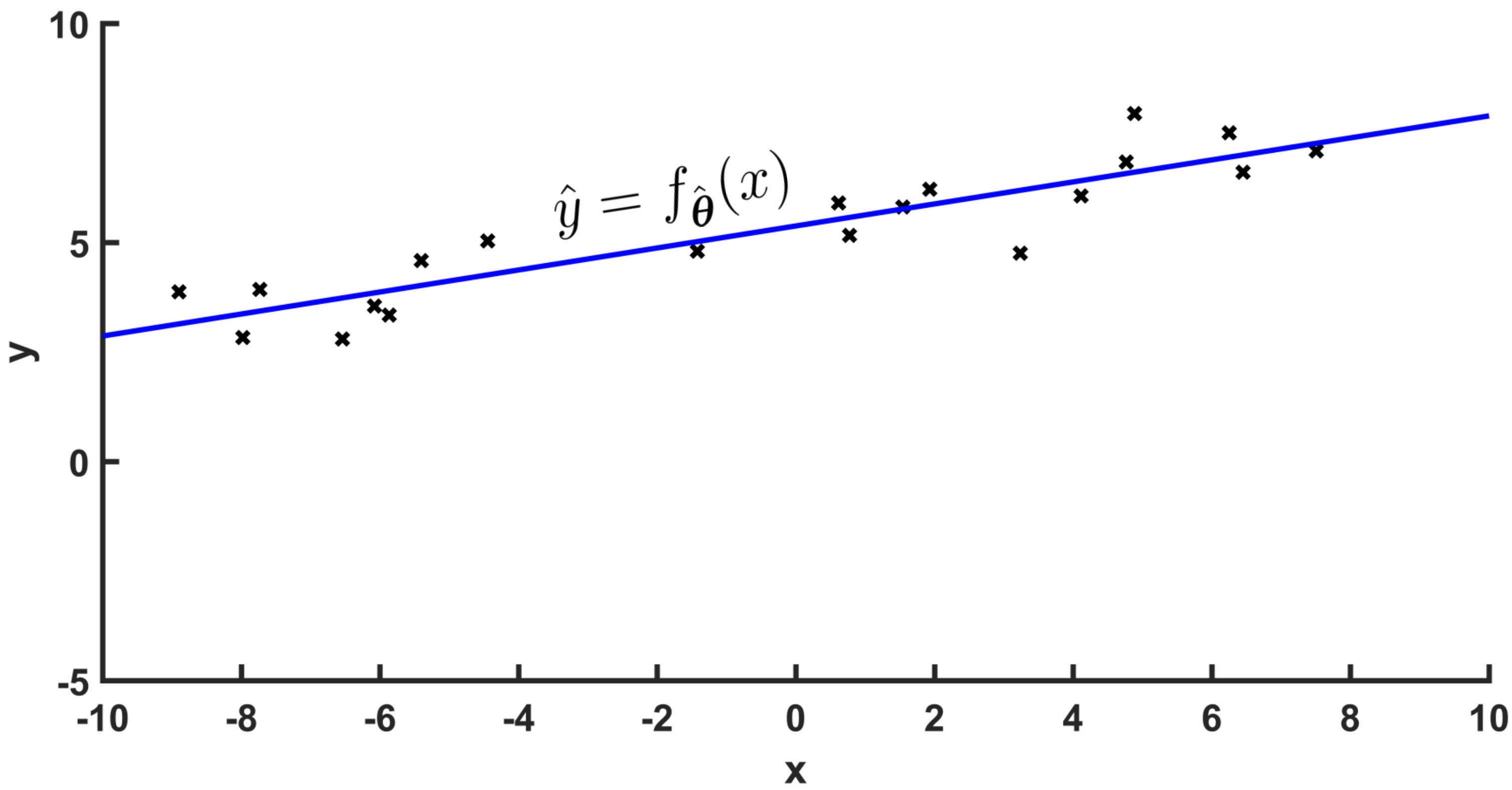


$$J(\boldsymbol{\theta}) = V(\boldsymbol{\theta}) + \lambda R(\boldsymbol{\theta})$$



Estimation du prédicteur

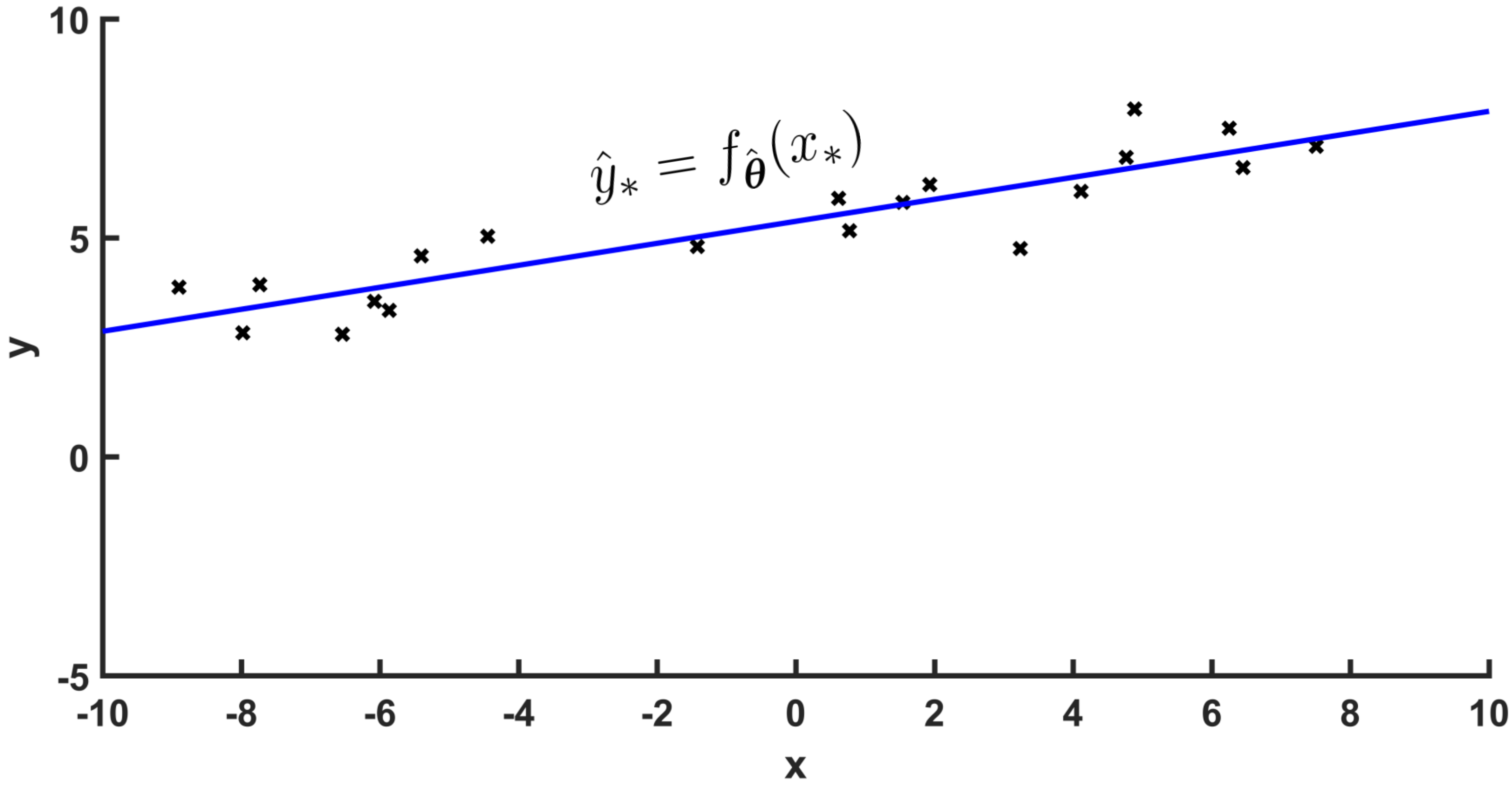
$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$



Maximum a posteriori

$$p(\boldsymbol{\theta}|X,Y) = \frac{p(Y|\boldsymbol{\theta},X)p(\boldsymbol{\theta})}{p(Y)}$$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid X, Y)$$



L'*a posteriori* comme fonction objective

(règle de Bayes)

$$\arg\max_{\boldsymbol{\theta}} \frac{p(Y \mid \boldsymbol{\theta}, X) p(\boldsymbol{\theta})}{p(Y)}$$

(transformation monotone)

$$\arg\max_{\boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta}, X) + \log p(\boldsymbol{\theta}) - \log p(Y)$$

(constante)

$$\arg\max_{\boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta}, X) + \log p(\boldsymbol{\theta})$$

(changement de signe)

$$\arg\min_{\boldsymbol{\theta}} -\log p(Y \mid \boldsymbol{\theta}, X) - \log p(\boldsymbol{\theta})$$

L'*a posteriori* comme fonction objective

$$\arg\min_{\boldsymbol{\theta}} \; -\log p(Y \mid \boldsymbol{\theta}, X) - \log p(\boldsymbol{\theta})$$

(exemple de régression linéaire)

$$\arg\min_{\boldsymbol{\theta}} \; -\log \prod_{i=1}^{N} \mathcal{N}(y_i \mid f_{\boldsymbol{\theta}(x_i)}, \sigma_y^2) - \log \prod_{j} \text{Laplace}(\theta_j \mid 0, b)$$

(simplification)

$$\arg\min_{\boldsymbol{\theta}} \; -\sum_{i=1}^{N} -\frac{(y_i - f_{\boldsymbol{\theta}}(x_i))^2}{2\sigma_y^2} - \sum_{j} -\frac{|\theta_j|}{b}$$

(simplification)

$$\arg\min_{\boldsymbol{\theta}} \; -\sum_{i=1}^{N} -\frac{(y_i - f_{\boldsymbol{\theta}}(x_i))^2}{2} - \sum_{j} -\frac{|\theta_j|}{b}$$

(multiplication par $\frac{2}{N}$ et changement de variable $\lambda = \frac{N}{2b}$)

$$\arg\min_{\boldsymbol{\theta}} \; \frac{1}{N} \sum_{i=1}^{N} (y_i - f_{\boldsymbol{\theta}}(x_i))^2 + \lambda \sum_{j} |\theta_j|$$

L'*a posteriori* comme fonction objective

$$\arg\min_{\boldsymbol{\theta}} \; -\log p(Y \mid \boldsymbol{\theta}, X) - \log p(\boldsymbol{\theta})$$

(exemple de régression linéaire)

$$\arg\min_{\boldsymbol{\theta}} \; -\log \prod_{i=1}^{N} \mathcal{N}(y_i \mid f_{\boldsymbol{\theta}(x_i)}, \sigma_y^2) - \log \prod_j \text{Laplace}(\theta_j \mid 0, b)$$

(simplification)

$$\arg\min_{\boldsymbol{\theta}} \; -\sum_{i=1}^{N} -\frac{(y_i - f_{\boldsymbol{\theta}}(x_i))^2}{2\sigma_y^2} - \sum_j -\frac{|\theta_j|}{b}$$

(simplification)

$$\arg\min_{\boldsymbol{\theta}} \; -\sum_{i=1}^{N} -\frac{(y_i - f_{\boldsymbol{\theta}}(x_i))^2}{2} - \sum_j -\frac{|\theta_j|}{b}$$

(multiplication par $\frac{2}{N}$ et changement de variable $\lambda = \frac{N}{2b}$)

$$\boxed{\arg\min_{\boldsymbol{\theta}} \; \frac{1}{N} \sum_{i=1}^{N} (y_i - f_{\boldsymbol{\theta}}(x_i))^2 + \lambda \sum_j |\theta_j|}$$

Le processus Beta

Le processus Beta est une infinité de distributions Beta :

$$\pi_k \sim \lim_{K\to\infty} \text{Beta}\left(\alpha\frac{\gamma}{K}, \alpha(1-\frac{\gamma}{K})\right) \qquad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$$